МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 7** **Москва 2020**

ОБЩИЙ РАЗДЕЛ

УДК 001.102

Н.В. Максимов, А.А. Лебедев

О природе и определениях информации: физика и семантика*

Рассматриваются результаты анализа публикаций, посвященных определению роли и сущности информации в различных предметных областях. Выделение аналогий общих понятий позволяет найти подходы к универсальному их определению. Анализируются свойства двойственности и действенности информации в информационных взаимодействиях, а также вопрос семантики информации, в том числе для физических систем, не включающих субъекта как активного участника.

Ключевые слова: *информация, теория информации, информационные взаимодействия, семантика*

DOI: 10.36535/0548-0027-2020-07-1

ВВЕДЕНИЕ

Роль информации фундаментальна: обладая особыми свойствами, она обеспечивает, наряду с энергией и материей, *управляемое* развитие самой среды, в которой, в том числе, живет человек.

Информация присутствует везде. Но при этом она многолика и, как следствие, возможности ее использования ограничены. С точки зрения семантики и прагматики – то, что будет полезно в одной предметной области, в другой – будет «шумом». Но более существенными являются ограничения, связанные с формой. С одной стороны, информация – это данные (данность, объект, сигнал и т.п.) в совокупности с контекстом (целью, предметом рассмотрения и т.п.).

* Работа выполнена при поддержке Министерства науки и высшего образования РФ (проект государственного задания № 0723-2020-0036)

Соответственно, чтобы данные донесли изначальный смысл, они должны быть поняты, т.е., на стороне приемника должен быть тот же или близкий контекст. С другой стороны, информация неизбежно представлена (хранится, передается) тем или иным носителем. В гомогенной (однотипной) среде практически не возникает проблема выделения данных: для взаимодействующих сторон это «родной» носитель, одинаковым образом используемый и источником, и приемником информации. То же самое относится к передаче контекста. В случае же, когда информация передается между средами, имеющими разную природу, необходимы специальные (по отношению к приемнику) сенсоры/интерфейсы, способные воспринимать «инородный» носитель и его состояния. И необходимо сформировать контекст, обеспечивающий адекватное формирование смысла на воспринятых данных, т.е., если это человек, то его надо обучить тому, что знает приемник в предметной области, если автоматизированная система, то сформировать словарь данных и тезаурус, располагать методом обработки.

В этом смысле важно иметь адекватное представление об информации, ее *свойствах* и *форме существования* как *объекте*, действующем и используемом не только в сфере информационных технологий. Только в этом случае можно объективно исследовать её структуру и свойства, а самое главное – обоснованно определять условия и средства её *эффективного* использования.

С той поры, как К. Шеннон в рамках теории информации ввел меру информации, было дано много и очень разных определений информации, которые по большей части были конструктивны только в той предметной области, в рамках которой они определялись.

Но ни известные концепции, ни определения, ни многочисленные публикации не дают четкого и убедительного ответа на вопрос о форме и/или способе существования информации: это физическая субстанция или свойство[1], преходящее или постоянное взаимодействие[2]; единая субстанция, или разные в разных сферах и областях; как соотносятся информация и смысл?

В настоящей статье на основе материалов известных публикаций[3] и фундаментальных подходов предпринята попытка «собрать» обобщенное представление об информации: как физическом явлении, средстве коммуникации в различных сферах, средстве управления и развития в сложных (социальных, живых и кибернетических) системах.

ПОДХОДЫ И ОПРЕДЕЛЕНИЯ ИНФОРМАЦИИ

Рассматривая информацию с точки зрения природы среды, Л. Бриллюэн [3] подразделил ее на 1) «связанную информацию» – когда возможные исходы представлены микросостояниями физической системы; и 2) «свободную информацию», когда возможные исходы рассматриваются как абстрактные значения. Во втором случае надо уточнить, что это абстрагированные от конкретного физического значения состояния, по крайней мере до того, как возникнет потребность измерять, сохранять и передавать нечто об этом физическом состоянии, т.е. информацию.

Согласно [4] информация представляет собой качественную и количественную характеристику организованности отражения. Вообще информация – это как бы некоторая "сила", направленная против дезорганизации и хаоса: в этом смысле информация неотделима от структурности, организованности материальных систем.

В целом можно выделить три сферы, где используется и достаточно четко определено понятие информации:

- физическая (статистическая термодинамика, квантовая механика), где она фигурирует преимущественно как мера упорядоченности (негэнтропия);
- криптография и связь (теория информации К. Шеннона);
- управление в сложных развивающихся системах (Н. Винер, У.Р. Эшби), а также информационные коммуникации в обществе и науке, где информация – это объект обработки и средство управления, существующие всегда в форме, обеспечивающей ее хранение, поиск и преобразование.

Соответственно, определения в этих сферах принципиально (категориально) различны. Но они относятся к одному предмету – понятию «информация». Поэтому их нельзя позиционировать как альтернативные: по существу, они являются взаимодополнительными, отражающими разные аспекты, роли, формы и фазы «жизненного цикла» того, что называется «информация». В частности, «связанная информация» представляет фазу «экзистенциональную» – физическую форму ее существования (внутрисистемные взаимодействия). «Свободная информация» – фазу «применения» – привычные операции сбора, хранения, передачи информации, использование ее в управлении (внешние взаимодействия).

«Информация есть информация, а не материя и не энергия» [5]. Информация присуща и абстрактному миру, и всем физическим процессам и системам[4]. С другой стороны, необходимо понимать, что чем бы ни были «вещи» в своем мире вещей, они могут войти в мир коммуникаций и смыслов только посредст-

[1] Свойство вещи (или процесса) не есть субстанция, оно не существует в вещи, а «… проявляется в отношении к другим вещам» [1].

[2] Аналогичная ситуация отмечается и в квантовой механике, где, по словам Р. Пенроуза [2], физики похоже не очень то стремятся к полной ясности в вопросе выявления «реальных» состояний и объектов.

[3] В данной работе авторы не приводят отдельным пунктом обзор литературы: классические работы цитируются по тексту, обзор перспективных направлений, связанных с физическим подходом, будет дан во второй части статьи.

[4] Точка зрения, что информация присуща лишь самоорганизующимся системам (прежде всего биологическим и социальным) в современной литературе практически не обсуждается. И, как представляется, «функциональная» концепция является частным случаем "атрибутивной", если учитывать двойственную природу информации, обладающей свойствами как объекта, так и процесса.

вом имен, качеств, признаков, т.е. посредством сообщений о своих внутренних и внешних отношениях и взаимодействиях [6].

Понятие информации, так или иначе, с одной стороны, связано с понятиями неопределенности, разнообразия, неоднородности, характеризующими состояние, а с другой – с целенаправленными изменениями, отражением, выбором, управлением. Помня, что все сущее имеет физическую форму, рассмотрим содержание одного из наиболее цитируемых, предельно обобщенное, концептуальное (и по отношению к разным предметным областям, и вследствие высокого уровня использованных понятий) определение: «Информация – это отраженное разнообразие» [7].

Существо концептуального определения информации

Представленное определение можно перефразировать следующим образом: "информация" – это устойчивое, стабильное *нечто* (*макрообъект*), называемое "отражение", получаемое в результате (*преходящего*) *действия* на другое *нечто*, называемое "разнообразие". Или, "информация" – это макрообъект (имеющий роль *образ*а), который появляется в результате *взаимодействия* макрообъекта "разнообразие" (имеющего роль *оригинала*) с отображающим объектом (имеющим роль *преобразователя*).

Таким образом, основное фундаментальное понятие "отражение" можно интерпретировать следующим образом: в результате взаимодействия отображающий объект *изменяется*[5] – на нем "остаются следы", по которым и можно судить о некоторых свойствах оригинала. Отсюда следует, что "отражающий" объект должен иметь структуру – устойчивые связи устойчивых частей[6]. И, в свою очередь, устойчивый характер изменения части (при неизменном состоянии остальных частей) отражающего объекта позволяет выявлять наличие образа, а устойчивая связь характера изменений с характером оригинала, вызвавшего изменения, – устанавливать, как факт, соответствие образа с оригиналом.

Другое следствие – это требование "совместимости" – объекты («отражаемый» и «отражающий») должны иметь общее средство взаимодействия (переноса), и взаимодействие должно осуществляться достаточно интенсивно, чтобы превысить порог, необходимый для заметного изменения состояния отражающего объекта.

Таким образом, здесь изменение отражающего объекта является способом отображения оригинала, причем осуществляющимся только при условии «совместимости». Существенно, что свойства оригинала в этом изменении представлены в *преобразованном* виде, через собственные характеристики отображающего объекта [8]. Очевидно, что "отражение" представляет собой "неактивное" преобразование, не добавляющее свойств и, соответственно, разнообразия (преобразователь в конкретный момент времени может находиться только в одном состоянии).

Другое представленное в определении [7] фундаментальное понятие "разнообразие" применительно к множеству может рассматриваться, согласно [9], как: 1) число различимых элементов множества (отдельных, но не обязательно разных); 2) число признаков (в общем случае множество или структура свойств – мерономия), принимающих значение 0 или 1 (или Log_n, если $n>2$ значений), необходимых для идентификации (различения[7]) *элементов множества*. Во втором случае мы имеем уже не однородное множество, а композицию, включающую два множества (пространства), между элементами которых установлено соответствие (на которых задано взаимное отображение). Здесь надо отметить, что утверждение «установлено соответствие» (как и выделение самих множеств из универсума) уже означает, что соответствие неслучайно, т.е. можно сказать, что оно имеет некоторый конкретный *смысл*. Так, информация – это разнообразие, которое один объект содержит в себе о разнообразии другого объекта.

Шенноновская информация

К. Шеннон рассматривает информацию как то, что устраняет неопределенность и, соответственно, измеряет ее количеством неопределенности, которую она устраняет. При этом мера, построенная на величине снятой неопределенности, показывает *прирост* информации от получения сообщения, а не абсолютное ее количество до или после этого.

Кроме того, существенную роль имеют принятые при этом допущения: 1) сумма вероятностей исходов должна равняться *1,* т.е. исходы образуют полную группу событий; 2) в случае источника с несколькими множествами вероятностей матрица переходных вероятностей должна быть марковской, т.е. вероятность перехода зависит только от текущего состояния; 3) энтропия определяется усреднением при условии *установившегося равновесия*. Таким образом, в этом случае объектом являются замкнутые системы, а информационные взаимодействия реализуются взаимно однозначными преобразованиями – кодированием-декодированием последовательности данных. Следовательно, шенноновская информация никак логически не связана со смыслом данных (это только передача сигнала – без контекста его появления или использования).

Информация в физической реальности

Физическая точка зрения отражена следующим фундаментальным определением: "Информация – это мера[8] неоднородности в распределении энергии (или вещества) в пространстве и во времени. Информация существует постольку поскольку существуют материальные тела[9] и, следовательно, созданные ими неоднородности" [12].

---

[5] "Изменяется текущее состояние" означает, что до взаимодействия объект стабильно находился в некотором другом устойчивом состоянии.

[6] В качестве "части" могут выступать как материальные "фрагменты" (элементы и связи), так и абстрактные сущности/отношения, обеспечивающие предметный анализ.

[7] "Неоднородность – иное выражение, вид разнообразия" [10].

[8] По нашему мнению, это еще и *образ.*

[9] Хранение и передача информации осуществляются физическими носителями. На микроуровне для этого надо использовать разного типа частицы: для хранения информации целесообразны ферми-частицы (стационарные состояния систем ферми-частиц устойчивы), а для переда-

Различимой неоднородностью в физике элементарных частиц может являться сама частица, ее состояние или связь с другой частицей. Исходя из этого в [13] представлены оценки объема информации для разных объектов – от фундаментальных частиц, молекул, черных дыр и до Вселенной. «Запутанность» может рассматриваться как элемент информации (где квант информационного взаимодействия – информация связи): неопределенность сцепленных состояний кубитов меньше суммы их неопределенностей невзаимодействующих состояний (по состоянию одного кванта можно определить состояние другого). Квантовая механика постулирует, что один бит соответствует элементарной физической системе, причем под элементарной системой понимается такая, которой соответствует истинностное значение одного суждения (а не элементарная частица или ее состояние!) [14].

Информация, понимаемая как устойчивые (определенное время) неоднородности произвольной физической природы [13], – это не собственно части среды, находящиеся в различающихся состояниях, а характер, *образ различий*.

Неоднородность физической реальности является следствием ее стохастической иррегулярной (индетерминированной) природы – это перманентные взаимодействия, реализующие следствия принципа неопределенности Гейзенберга. При этом (вследствие сложившегося[10] разнообразия видов физического взаимодействия, спутанности, принципиального стремления к состоянию динамического равновесия) возможно появление устойчивых "схем" распределения составляющих, различаемых извне по некоторым признакам – собственным свойствам, состояниям.

Постоянное изменение неоднородной и неравновесной среды может привести к появлению упорядоченных (структурированных) макрообъектов и становлению форм поведения, что, в свою очередь, за счет взаимодействия с окружением, обеспечивает устойчивость самого объекта или направления его развития. Объяснение этого связано с понятием «детерминированный хаос»: случайность в такого рода системах хотя и обязательно имеет место, но ограничена[11]. Так, предполагается, что имеется (появляется изначально или формируется в процессе вследствие законов взаимодействия) некоторое преимущественное направление развития процесса. При этом элемент случайности обеспечивает возможность появления нового, которое, так или иначе, приводит к нарушению устоявшейся системы, её достраиванию или выходу за собственные пределы [16]. Такая возможность предопределена динамическим характером предметных областей: развивающиеся системы – это не абсолютно стабильные, неоднородные среды, где «целью» является достижение устойчивости, эффективности роста системы или её взаимодействия с окружающей средой, а «источником» развития - состояние динамического хаоса [17, 18].

В частности, квантовые системы могут коррелировать по принципу непричинной, несиловой связи (*парадокс Эйнштейна-Подольского-Розена)*, суть которого в том, что пространственно разнесенные части одной системы мгновенно приобретают информацию друг о друге [19]. В работах В.Ю. Татура [20] данный феномен проявляется в так называемом пространстве *"слабой метрики"*, соответствующей такой форме материи, для которой квантовые объекты описываются как единые и неделимые, как целое. Для пространства слабой метрики характерны квантовые аналоги пространства – времени. Поэтому его можно представить как многоуровневое параметрическое пространство. Между уровнями происходит движение квантового перехода с изменением энергетического состояния.

Каждый из основных уровней имеет множество подуровней. Каждый уровень организует в метрическом пространстве макроквантовую систему, объект (элементарная частица, клетка, организм, биосфера, галактика и т.д.) как единое и неделимое целое. Согласно В.Ю. Татуру [20] пространство со слабой метрикой влияет на движение элементарных частиц посредством «аксионов», которые, поскольку являются лишь этапом в становлении проявленного мира, взаимодействуют со слабой метрикой сильнее, чем элементарные частицы.

В повседневной жизни мы почти не замечаем квантовых явлений. Возможность наблюдать классическую (детерминированную) реальность можно связать с тем, что при взаимодействии квантового объекта с окружением происходит многократное запутывание его состояний с окружением [21]. В результате на окружении возникают множественные «отпечатки» только особых состояний объекта из всех возможных, которые В. Зурек назвал «индикаторными». Среда выступает инструментом отбора возможных состояний, усиливая процесс «выживания»[12] индикаторных состояний, а множественность (избыточность) «отражений» объекта в окружении позволяет различным наблюдателям извлекать из окружения одни и те же данные об объекте.

Таким образом, для физической реальности информация – это неоднородность распределения неоднородностей (или – данные в контексте вектора развития).

Информация в физике

Неопределённость состояний, в которых может находиться физическая система, в статистической физике характеризуется энтропией. Неопределённость конкретного состояния, точнее неопределённость изучаемой системы, находящейся в некотором состоянии, характеризуется *информационной энтропией*. Если *наблюдаемая*[13] фиксирована, то можно

чи – бозе-частицы (имеют много состояний и не могут находиться в состоянии покоя) [11].

[10] «Изотропию и однородность физического пространства – его евклидовость (псевдоевклидовость) – можно объяснить его простотой. Это единственное максимально симметричное пространство с нулевой кривизной» [15].

[11] Вследствие принципа минимального действия, стремления к устойчивому состоянию.

[12] Поэтому данная интерпретация получила название «квантовый дарвинизм».

[13] Под «наблюдаемой» в квантовой физике понимают, с некоторыми оговорками, физическую величину, измеряе-

говорить о неопределённости ее состояния, т.е. энтропия является собственным параметром системы, сама по себе характеризующая ее неупорядоченность. Информационная же энтропия характеризует неупорядоченность образа системы. Соответственно, неопределенность эксперимента складывается из этих двух составляющих. Таким образом, наблюдаемая физическая система, с которой взаимодействует наблюдатель, имеет и внешнюю, и внутреннюю неопределённость, которые в сумме составляют полную неопределённость системы. Соответственно, информация определяется на наблюдаемых состояниях системы посредством измеряемых величин [13].

Появление наблюдаемых может зависеть или нет от некоторого субъекта (плана эксперимента, потребностей управления и т.п.). В первом случае – это, по терминологии Л. Бриллюэна, «свободная информация»; во втором – «связанная информация», которая бывает, в частности, в естественных природных процессах, сопровождающихся каким-либо явлением (появлением/исчезновением материи/энергии). Действительно, для объективной реальности законы физики предопределяют, что неупругие физические взаимодействия порождают дополнительные состояния или объекты (не существовавшие в системе до взаимодействия) [22], которые могут использоваться для получения сведений о свойствах (процесса или самих взаимодействующих) – т.е. обеспечивают своим существованием проведение измерения.

В квантовых теориях при постановке эксперимента могут выделяться (исходя из физической модели) отдельные наблюдаемые, измерение значений которых (состояния *реальных* объектов/процессов) не может быть произведено одновременно (т.е. соответствующие выбранным наблюдаемым операторы не коммутируют). Амплитуду и фазу волновой функции измерить (непосредственно) тоже невозможно: "Мы ищем непрерывную волну, а получаем дискретные частицы. …. При этом волновое уравнение не рассказывает всей истории. Мы должны исследовать условия наблюдения, граничные условия и обстоятельства, при которых волны могут быть испущены и поглощены" [1]. Необратимость, в частности, может быть обусловлена несимметричностью граничных условий: источник и измеритель находятся (совмещены) на границе и граничные условия (последовательность изменения энергетических состояний взаимодействующих элементов) несимметричны, в том числе и во времени.

«Измерения производятся не в абстрактном гильбертовом пространстве. Экспериментаторы прекрасно знают какую наблюдаемую они хотят измерить, и соответственно выбирают взаимодействие, которое и определяет предпочтительный базис» [23].

Также нельзя провести наблюдение без возмущения объекта. Любое, даже пассивное наблюдение вносит возмущение, поскольку происходит поглощение по крайней мере нескольких квантов. Взаимодействие конечно и определено Планковской константой.

Таким образом (условно-)целостную картину реальности мы "складываем" (что тоже привносит некоторую неопределенность) из аспектных представлений – результатов измерений отдельных наблюдаемых. Принципиальная возможность существования "дополнительности" (в эксперименте и не только) следует из того, что "разрезание" целостной наблюдаемой системы, разрушая ее структуру (связи), порождает "валентности" ("дырки"/"избыток"), которые обеспечат "совместимость" при последующем "сшивании" ее частей. Но, соответственно, дополнительность имеет своим естественным следствием неопределенность.

Эксперименты – единственное средство познания, доступное человеку: реальное значение имеют только наши ощущения. По мнению М. Планка, мы наблюдаем некоторые регулярности в экспериментах и формулируем эмпирические законы, позже увязываем их с теми или иными теоретическими моделями, которые по существу являются результатом вымысла и достоверно совпадают с реальностью только в точках измерения с точностью погрешности инструмента. Плюс к тому, научная теория состоит из двух частей:

1) системы абстракций, представляющей зависимости (законы) для величин, выбранных в качестве основания теории;

2) правил (а также приборов и интерфейсов) соотнесения этих величин и зависимостей с физической реальностью, обеспечивающих измерения.

Отсюда в совокупной системе «теория – физическая реальность» есть три вида неопределенности: 1) аксиологическая неопределенность системы абстракций; 2) субъективная (импликативная) неопределенность эксперимента/измерений; 3) объектная неопределенность состояний физической реальности.

Теория (закон) описывает систему всегда в определенном аспекте, тогда как в действительности могут быть выделены и другие аспекты. Но при этом в отношении эквивалентных теорий всегда можно выделить класс содержательных утверждений, которые сохраняются во всех этих теориях. Утверждение Л. Бриллюэна "теория заключает в себе фундаментальную неопределенность" является следствием теоремы Геделя и принципа дополнительности как «механизма». Теория создает негэнтропию. "Негэнтропия структуры (теории, машины) не равна сумме негэнтропий ее компонентов" [1], поскольку уже само соединение частей, выполняемое для реализации определенной функции, является ограничителем разнообразия возможных состояний каждого из компонентов. Информация в этом случае – это компоненты в контексте конструкции (схемы соединения частей).

Таким образом, информация в физике – это измеряемые наблюдаемые в контексте теории – совокупности параметров и уравнений. Основные законы ограничивают многообразие систем, определяют допустимые варианты их построения и функционирования [11].

Информация в управлении и коммуникациях

В обществе, науке и управлении информация – это основной объект обработки и средство управления. Это не только широко используемый термин, но и хорошо определенное понятие, в том числе на

мую в эксперименте. Более точно – оператор, описывающий состояние физической системы.

уровне международных и российских стандартов, где даются следующие определения информации:

- это знания о предметах, фактах, идеях и т.д., которыми могут обмениваться люди в рамках конкретного контекста [24];
- это знания относительно фактов, событий, вещей, идей и понятий, которые в определённом контексте имеют конкретный смысл [25];
- это сведения, воспринимаемые человеком и (или) специальными устройствами как отражение фактов материального или духовного мира в процессе коммуникации [26].

В [27] уточняется, что информация должна иметь такую форму представления, которая обеспечит коммуникацию (хранение, поиск, передачу); информация – это, в первую очередь, интерпретация (действенность) – смысл этого представления.

Приведенные определения вполне соответствуют и изначальному смыслу латинского слова *informātiō* – «разъяснение, представление, понятие о чём-либо», и его глагольной форме – *informare* «придавать вид, форму, обучать; мыслить, воображать».

Таким образом, информация – это **данные** (материальная форма характеристики объекта – его образа, содержание) **в контексте** «чего-то». Такой контекст может быть задан методом обработки или применения; сопоставлением, например, с назначением, а также с другими конкретными данными. Причем эти данные – сведения (т.е. данные, собранные не случайным образом) существуют независимо от формы их представления

ИНФОРМАЦИОННЫЕ ВЗАИМОДЕЙСТВИЯ

Ранее мы показали, что информация связана с устойчивыми неоднородностями, которые являются следствием перманентных взаимодействий в природе (точнее, информация – это образ различий). Теперь рассмотрим характер и особенности взаимодействий, связанных с информацией.

Механизмы (характер) взаимодействия

На уровне физики микромира, изучающей фундаментальные законы и составляющие Вселенной, никаких взаимодействий, кроме известных четырех фундаментальных видов[14], не требуется. Но следует особо отметить, что в этом случае рассматриваются *элементарные* частицы, взаимодействующие по *элементарной* схеме (минимальное число участников – 2, и, обычно, один акт взаимодействия), а «элементарность» определяет не только фундаментальность (как «первооснову» и неделимость), но и однородность (одинаковость, а не равномерность и не равновесность) операционного пространства, позволяющего *статистически* не различать наблюдаемые экземпляры. На этом уровне в роли такого механизма взаимодействия выступают устойчивые соотношения свойств взаимодействующих объектов, отражаемые законами физики (константы, уровни, переходы и т.п.), в том числе (стабильно) допускающие при микровоздействии вызывать макроизменения того объекта, на который оказывается воздействие.

В то же время на этом уровне возможна и самоорганизация, когда при определенных условиях в открытых системах возникают устойчивые пространственные или временные структуры. Показательным примером, приведенным в [28], является процесс перехода излучения в лазере от спонтанного в состояние когерентного. Здесь *циклическая причинность* возбуждения атомов приводит к усилению амплитуды волны и, при достижении определенной (достаточно большой) амплитуды, волна *становится когерентной.* Так, возникающая волна, «подчиняющая» атомы, служит управляющим параметром[15], но и она сама порождается совместными их действиями (атомы «кооперативно» путем обмена информацией «создают» среду, из которой они получают информацию как действовать когерентно). Но необходимо отметить, что такое взаимодействие возможно при довольно ограниченных условиях: энергиях, веществах и временах.

В природе наблюдаются и неодноактные взаимодействия неэлементарных макрообъектов, которые уже не удовлетворяют требованию «однородности» и результат которых определяется не только свойствами взаимодействующих объектов, но и другими факторами: начальными условиями, неоднородностью объекта или его окружения и т.д. Фактически взаимодействие доопределяется макросвойствами[16] и, с точки зрения теории управления, такие процессы будут относиться к классу параметрических.

На уровне физических взаимодействий в результате эволюции возможно возникновение и менее ресурсоемких схем – взаимодействие посредством *образов* – объектов, некоторым образом сопряженных с объектами среды (*оригиналами*), процесс взаимодействия которых будет менее затратным. Образы и оригиналы могут принадлежать одному пространству – в этом случае взаимодействие сводится к известным физическим формам, или разным, и в этом случае взаимодействие будет опосредованным.

Биосферный уровень уже характеризуется явным использованием специализированных структур, обеспечивающих наследование и адаптацию свойств (формирование вида) и, в итоге, более эффективное поддержание устойчивости за счет це-

[14] В настоящее время известно четыре *фундаментальных* вида физических взаимодействий: слабое, сильное, электромагнитное, гравитационное. Все остальные виды взаимодействия, в том числе и людей, имеют в своей основе перечисленные. Наблюдаемая действительность показывает, что все функции сознания (получение, интерпретация, генерация информации пр.) реализуются физическими и химическими процессами.

[15] В синергетической теории Г. Хакена [28] это «параметр порядка», по аналогии с теорией фазовых переходов.

[16] Существование в природе фундаментальных частиц и взаимодействий также определяется макросвойствами – наблюдаемой устойчивостью форм объектов и превращений окружающей действительности. Или, согласно Ч. Пирсу [29] – это своеобразные «привычки» природы – космологические регулярности. Они и позволяют нам «… иметь поведенческую установку действовать неким определенным образом, и предписание для данного действия» [30].

ленаправленного воспроизводства предопределенных свойств «по образу»[17].

Именно такие реальные объекты – образы, отражающие свойства (но не обладающие ими) некоторого оригинала и использующиеся для создания нового объекта (оригинала в той же действительности), обычно и называют информацией[18].

Собственно информационные взаимодействия – это не взаимодействия неких «информаций», а взаимодействия физические, схема которых (соотношение составляющих, порядок, условия и т.п.) имеет некоторые специфические особенности выполнения или применения.

Условия информационных взаимодействий

Все используемые и изучаемые человеком объекты и их взаимодействия, так или иначе, связаны с физическим миром.

Это относится и к абстрактным[19] объектам (которые существуют в сознании «физического» человека), поскольку их взаимодействие с окружающим миром возможно, только если представить их материальными средствами: языком коммуникаций или воплощением в изделиях.

В физической реальности «причиной» события (преобразование объекта или его состояния) является некоторая «сила», действие одной части физической системы на другую. В мире идей такой причиной будет выделяемое субъектом соотношение частей, или их состояний в разные моменты времени. Причем, для «проявления» (объективизации) результата этого соотношения необходим отдельный ***вещественный*** компонент (возможно другой системы) – *приемник*, реагирующий на *различие* или *изменение*. При этом различие уже есть *невещественный* феномен (связанный с негэнтропией, а не с энергией).

Без дифференциации частей не может быть дифференциации событий или функций[20]. Сложные процессы могут возникать только благодаря взаимодействию частей (частиц) [6]. Это относится не только к физической, но и к ментальной сфере: «Вещи, не имеющие между собой ничего общего, не могут быть и познаваемы одна через другую» [31]. Утверждение в равной мере относится и к физической, и к ментальной сфере, и является одним из основных обоснований для определения свойств интерфейса: взаимодействующие части должны иметь соответствующую общность.

Особенность информации в том, что она объединяет в себе «содержание» и «форму». Причем это явление *одновременное* и *однопричинное* [32]. Но «содержание» представляется одним языком (в физической основе которого знаковая система), обеспечивающим построение и использование смысловых, ценностных компонентов, а форма реализуется другим языком (знаковым представлением) – формированием физического описания. По существу, два типа описания, произведенные на языках, не имеющих между собой смысловых связей, объединяются в одной структуре. При этом содержание логически не зависит от его физического описания (что вполне соответствует принципу изофункционализма систем А. Тьюринга о возможности воспроизведения системы с данным набором функций на разных элементных базах). Это *принцип инвариантности информации по отношению к физическим свойствам ее носителя*: информация необходимо воплощена в определенном физическом носителе, но одна и та же информация может иметь разные по своим физическим свойствам носители. Отметим, что кодированные зависимости становятся средствами и элементами самоорганизации, над которыми, в свою очередь, могут надстраиваться кодовые зависимости более высокого уровня.

ИНФОРМАЦИОННЫЕ ОБЪЕКТЫ, ИНФОРМАЦИОННЫЕ ВЗАИМОДЕЙСТВИЯ И ИНТЕРФЕЙСЫ

Предметной областью (ПрО) будем считать совокупность объектов (и/или их состояний) естественного или искусственного происхождения, либо существующих в виде сложившегося в результате эволюции устойчивого естественного образования, либо выделяемых некоторым субъектом в соответствии с целями его деятельности. Очевидно, что границы таких ПрО будут всегда достаточно условными.

Информационными объектами[21] (ИнфОб) будем считать те, которые обладают свойством (способностью) отражать и, при соблюдении определенных условий, изменять состояние другого физического объекта или его взаимосвязи [33]. Информационные объекты для нас *представляют* «атомы» информации: это та форма существования информации, которая обеспечивает возможность её хранения и передачи в пространстве и во времени. Информационный объект, порожденный каким-либо другим объектом, существует независимо от него. Такой объект не тождественен ни объекту, который был причиной его появления, ни тому, на который он воздействует, и в то же время он сам может быть объектом информационного воздействия.

Информационной средой (ИСр) предметной области будем считать множество информационных объектов, специально создаваемых либо выделяемых в этой или какой-либо другой ПрО.

Выделение во множестве объектов двух, в общем случае, пересекающихся подмножеств (ПрО и ИСр) *в принципе* условно. Объект, являющийся «информационным» (играющий роль) для одной предметной области, необязательно будет таковым для другой: для

[17] Механизм генетического воспроизводства предполагает устойчивость алфавита и постоянство механизма кодирования/декодирования, реализующего синтез объекта, подобного оригиналу по его образу (коду).

[18] Здесь следует подчеркнуть, что такая схема взаимодействия и взаимообусловленности *не антропоцентрична*, она не требует наличия сознания человека: для этого достаточно условия неравновесности среды.

[19] Отметим, что единственная известная на сегодня «машина», реализующая преобразование (взаимодействие) абстрактных объектов в реальные – это человек.

[20] Сюда же относится и квантование пространства – времени, что есть проявление динамичности микромира.

[21] Термин «объект» здесь используется для того, чтобы не акцентировать внимание на его природе – физической или абстрактной.

различных состояний системы (как и для различных предметных областей) эти разбиения будут разными.

В общем случае информационные взаимодействия (ИнфВ) – это такие процессы, в которых в качестве операционных используются объекты-образы. В отличие от физических взаимодействий, непосредственно происходящих между отдельными телами или силам (оригиналами) в конкретный момент времени, в информационных взаимодействиях будут участвовать и их образы. Причем такие объекты-образы способны вступать между собой во взаимодействие уже без ограничений, присущих оригиналам [34], что позволяет для развития оригиналов использовать эволюцию их образов. Информационные объекты обладают потенциальной возможностью по изменению состояния данной ПрО, или, другими словами – имеют свойство «действенности» в определенных условиях. Существенно, что свойство действенности (как и событие его появления или отнесения к информационным) имеет стохастическую природу, но проявиться может только при *взаимодействии* информации с соответствующей ПрО. Заметим также, что с субъективной точки зрения *информационные* взаимодействия отличаются от физических еще и тем, что назначением информационного объекта может быть не его собственное существование, а изменение состояние взаимодействующей с ним части ПрО.

Процессы такого типа могут (а для случая «свободной информации» – должны) использовать память – внешнюю по отношению к объектам-оригиналам операционную среду объектов-образов. Для реализации взаимодействия среда должна иметь механизмы[22] сопряжения – **интерфейсы** взаимодействующих объектов. Причем интерфейсы могут быть разной сложности: элементы, агрегаты, связь/отношение, структура/функция, система (функционал агрегатов и функций). Для восприятия управляющего сигнала датчик (сенсор) должен работать в соответствующем диапазоне частот и энергий; для декодирования шифрованного сообщения необходим ключ кодирования; для восприятия информации и синтеза на ее основе новых знаний тезаурусы источника и приемника должны иметь соответствующую общность.

Информационное взаимодействие – это многокомпонентный процесс, предполагающий *физические источник/приемник* и *каналы* передачи сообщения, *синтаксис* представления сообщения для его передачи и последующей *семантической* интерпретации в соответствии с тезаурусом ПрО, и, наконец, определение *прагматической* ценности в соответствии с конкретными целями и критериями субъекта. Кроме того, возможны мотивационные и сигнальные компоненты: приемник должен ожидать сообщение и уметь выделить его из потока других сообщений.

Информационное взаимодействие отличает нелинейность (необратимость), обусловленная дискретным характером процесса отображения сообщения из одного пространства в другое, когда происходит (производится) «редукция» свойств оригинала при генерации образа, происходящей при выборе определенного отображения (преобразования, системы кодирования).

Такое преобразование необходимо приводит к потере в образе разнообразия *возможных* состояний ПрО[23]. Даже при отображении ПрО «самой в себя» происходит потеря разнообразия, в данном случае – временных состояний, так как образ будет фиксировать свое состояние только на момент выбора. Но необходимо отметить, что это не уменьшение количества объектов/связей и их состояний (их стало даже больше за счет объектов-образов), а ограничение их множества, что будет использоваться при выборе[24] конкретно обусловленной выборки (в отличие от случайной, когда результат полезен только для сравнения с результатами упорядоченных, целесообразных действий).

С энергоматериальной точки зрения характерным свойством информационного взаимодействия является то, что для конкретной предметной области информационные объекты обладают способностью «переключающего воздействия». Для такого *информационного взаимодействия*[25] характерно:

1) энергетика порождения информации (точнее, отдельного объекта ИСр, отражающего свойства ПрО) настолько слаба, что соответствующее изменение состояния ИСр не влечет существенного изменения состояния ПрО;

2) предметная область имеет *точки бифуркации*, где сравнительно малая энергия информационного объекта влечет существенные изменения состояния объекта предметной области. Такая точка может быть следствием «естественной» или искусственно создаваемой неустойчивости предметной области, например, через такое устройство переключения, как триггер.

В общем случае это диссипативная система: для генерации результата используются дополнительные внешние энергоматериальные ресурсы[26].

---

[22] Для уровня фундаментальных взаимодействий – это законы Вселенной; для биосферы – генетические механизмы; для техносферы – системы управления.

[23] Отметим, что эти потери разнообразия *состояний* оригинала, отражаемых («свертываемых») в отдельном образе, для системы в целом могут быть «компенсированы» за счет разнообразия ассоциированных объектов на уровне образов. И как следствие, система получает возможность эволюционировать (и даже энергетически более экономно и эффектно) за счет взаимодействий образа, а не самого оригинала. «Мы плывем вверх по течению, борясь с огромным потоком дезорганизованности. … В этом мире наша первая обязанность состоит в том, чтобы устраивать произвольные островки порядка и системы» [35].

[24] Понятие «выбор» уже предполагает наличие предпочтений, т.е. упорядоченность, «неравнопредпочтительность» элементов множества, на котором осуществляется выбор.

[25] Этот тип взаимодействия не является еще одним видом упомянутых ранее фундаментальных видов физического взаимодействия, а только отражает характер соотношения взаимодействующих сторон.

[26] За исключением случая «связанной» информации (согласно определения Л. Бриллюэна), когда в качестве примера можно рассматривать излучение («сопровождающего» продукта деления/синтеза вещества), «используемое» в качестве информации для восстановления в той или иной форме (повторения в других формах и обстоятельствах)

Согласно [36] информация не тождественна информационному объекту. Если информационный объект ассоциируется с формой существования, то информация – с содержанием, т.е. действием, потенциально производимым этим объектом. С точки зрения физики процесса, ***информация*** – это ***информационный объект во взаимосвязи*** с теми объектами, с которыми он может «информационно» взаимодействовать, или иначе – информационный объект и некоторое предопределенное направление. Так информация в фазе стабильного состояния (хранимый или передаваемый информационный объект) обладает свойствами макрообъекта, а в фазе взаимодействия – волновыми. До взаимодействия – это некоторый целостный объект, во время взаимодействия – это макрообразование «квантов». Причем эти связанные микрообъекты[27] могут использоваться и в качестве элементов образов для других, существующих или будущих, объектов-оригиналов.

Отсюда следует, что информация как *суперпозиция возможных состояний информационного объекта* может быть представлена в *n* предметных областях и характеризоваться функцией:

$$\Lambda = c_1E_1 + c_2E_2 + ... + c_nE_n ,$$

где E_i – функция распределения[28] ИнфОб для *i-й* предметной области;

c_i – вероятностный [29]показатель отнесения ИнфОб к *i-й* предметной области.

Тогда *информационное взаимодействие* можно характеризовать функцией:

$$\Omega = c_1E_1\theta_1^T + c_2E_2\theta_2^T + ... + c_nE_n\theta_n^T ,$$

где θ_i – функция распределения для информационного объекта, с которым взаимодействует исходный ИнфОб, для *i-й* предметной области.

Последнее выражение отражает то, что информационное взаимодействие несимметрично (неассоциативно) и имеет «направленность», задаваемую для *воздействующего* информационного объекта вектором коэффициентов ***C,*** состоящим из элементов c_i. Именно при взаимодействии возникают неравновесные комбинации, создавая предпосылки для появления таких качеств, как ценность и новизна, которые станут характеристиками нового ИнфОб – результата взаимодействия.

Как мы отметили ранее, в процессе информационного взаимодействия появляется объект (образ), связанный с оригиналом логически. Причем исходный информационный объект, практически не меняя своего физического состояния, явно или неявно обретает новое свойство - «быть использованным в данной ПрО». Очевидно, что после этого результаты взаимодействия этого объекта с другими предметными областями не будут прежними. В итоге взаимодействия – выбора *i*-й ПрО в суперпозиции возможных состояний до момента взаимодействия, осуществляемого в среде некоторого *j*-го субъекта, происходит «редукция» функции распределения – из всех областей выбирается одна, которая становится «основной». Именно выбор *приоритетного* направления изменяет функцию распределения для последующих взаимодействий, уменьшая вероятности использования этой информации во всех других ПрО, возможно, до незначимой величины. В результате этого информационного взаимодействия появляется (или изменяется) *личное знание j*-го процессора в *i*-й ПрО – новый информационный объект, который, в свою очередь, может стать доступным для взаимодействия, в том числе, вне операционной среды *j*-го процессора. При этом тот же ИнфОб может использоваться другими субъектами и в других областях, взаимодействие с которыми будет порождать знание, в общем случае не тождественное *j*-му.

Отсюда следует важнейшее, характерное для систем, свойство информации – э*мерджентность*. Это означает, что информационная среда сама обладает способностью образования новых свойств, в том числе таких, которые позволяют обнаруживать новые свойства отражаемой ПрО. Причем, системность – это "информационное" (а не собственное) свойство системы, которая рассматривается (дополняется) в контексте предметной области и ее свойств.

СМЫСЛ, ЗНАЧИМОСТЬ, ЦЕННОСТЬ

Обычно смысл понимается как нечто, содержащееся в сообщении (тексте, сигнале), то, по чему мы можем судить об обозначаемом (что это, как устроено). Смысл – это *существо* (архетип) – образ, построенный (на сообщении) в соответствии с некоторой схемой (точкой зрения, шаблоном), отражающей интересы приемника. Не совсем корректно говорить «сообщение *имеет* смысл»: за редким исключением сообщение (если оно не тривиальное) «имеет» много смыслов. Сообщение в общем случае – это информация, результат взаимодействия «приемника» (точнее его информационного образа), который зависит не только (и даже не столько) от содержания сообщения, сколько от приемника (точнее, информационного образа, контекста), а также от метода, реализующего данное информационное взаимодействие. И если источник «заложил» определенный смысл в передаваемое сообщение, то не обязательно приемник получит именно этот смысл, так как приемник,

---

процесса, которого уже нет (модель прошлого), или которое, возможно, случится (прогнозирование будущего).

[27] Для знаковой (текстовой) формы информации такими квантами являются понятия, обозначаемые в тексте терминами языка, которые могут использоваться в любых других текстах.

[28] Для текстовой (точнее, дескриптивной) формы представления информации, реализующей координатный метод идентификации содержания, пространство (первичные координаты) задается на дискретном множестве терминов, представляемых целостными понятийными конструкциями слов и их комбинаций (словосочетаниями, фразами, полными текстами). Функция распределения в этом случае может быть задана с помощью матриц типа «термин-документ» для соответствующих ПрО.

[29] В общем случае c_i - комплексное число, так как предметные области представлены в не одномерном пространстве и их выделение определяется известным знанием, т.е. необходимо предполагать, что существуют и другие «части» ПрО, «мнимые» для текущего состояния.

по существу, пытается «отыскать» в сообщении не то, что туда положили (скорее всего, этот смысл ему неизвестен, по крайней мере, в полном объеме), а то, что *ему нужно*, что он знает (в том числе и о том, чего не знает).

Причем смысл может иметь как содержательный, так и идентифицирующий характер. В первом случае – это представление существа в некотором аспекте, во втором – обозначение, называние этого аспекта (например, «Координаты в смысле географии»). Таким образом смысл – это формируемое информационным взаимодействием новое или конкретно ассоциированное знание приемника – редукция информации (представляемой как суперпозиция возможных состояний информационного объекта) по отношению к контексту (знанию, тезаурусу) приемника. С технологической точки зрения [37] этому соответствует операция *аспектная проекция* – отображение онтологического представления содержания на представление контекста (онтология обстоятельств конкретной ситуации, предмета, точки зрения) приемника.

Так, сигналу смысл можно только *приписать,* и только в том случае, если принимается во внимание отклик – изменение состояния приемника[30]. Получение сигнала приемником означает, согласно [28, с.35], что с его помощью задается вектор и параметры среды (аттракторы, случайные флуктуации) и со временем вектор стремится выйти на аттрактор.

Отсюда значимость (ценность) сигнала (информации в сообщении) может определяться через значимость аттрактора, к которому стремится система после получения сигнала.

ВЫВОДЫ

1. Информационные объекты и взаимодействия являются объективной и закономерной действительностью, существующей наряду с фундаментальными элементарными частицами и их взаимодействиями. Более того, несмотря на то, что информационные объекты и взаимодействия реализуются, в конце концов, в той же «элементной базе» фундаментальных частиц и взаимодействий, они не сводятся фундаментальным частицам, выступая как надстройка в виде законов, констант и параметров порядка.

2. Информация имеет двойственность состояния. Информационный объект (части или состояния некоторого носителя) до взаимодействия – это некоторое цельное неделимое образование, во время взаимодействия – это макрообразование, структура «квантов», которые также могут быть составляющими элементами или комбинациями и других, существующих или гипотетических, объектов. Информация в фазе стабильного состояния (хранимый или передаваемый информационный объект) обладает свойствами макрообъекта, а в фазе взаимодействия – волновыми свойствами.

3. Информация имеет двойственную природу своего проявления: с одной стороны – это объект, а с другой стороны – это «действенность», приводящая к изменениям в результате взаимодействия с другими объектами. Здесь мы имеем динамику формы – от свойства элемента к элементу, обладающему свойством. В [38] представлен процесс преобразования тепловой энергии в механическую за счет информации, которая «передвигает» непроницаемые стенки внутри физической системы.

Особая разновидность информационного взаимодействия – «переключающее воздействие», для которого характерно, что энергетика порождения информации настолько мала, что соответствующее изменение состояния информационной среды не влечет существенного изменения состояния предметной области. Соответственно, предметная область имеет естественные или искусственные точки бифуркации, где сравнительно слабая энергия информационного объекта влечет существенные изменения состояния объекта предметной области.

Свойство «действенности» может проявляться только при *взаимодействии*. Для реализации взаимодействия среда должна иметь механизмы сопряжения – интерфейсы, т.е. взаимодействующие объекты должны иметь соответствующую общность**.**

4. Информация неизбежно имеет семантическую природу: данные, сигналы, если они выступают в роли информации, *всегда связаны с контекстом* – знанием, целями, ситуацией приемника, в частности:

- информация физической реальности как неоднородность распределения неоднородностей – это данность в контексте вектора развития. Это не собственно части среды, находящиеся в различающихся состояниях, а характер, образ различий;
- информация в физике – это измеряемые наблюдаемые в контексте теории (совокупности параметров и уравнений);
- информация в коммуникациях – это разнообразие, которое один объект содержит в себе о разнообразии другого объекта.

В целом, "Информация – любое небезразличное различие" [6].

Пользуясь сложившейся в информатике терминологией, информацию можно определить как совокупность {<*данные & метаданные*> & *метаинформация*>}, где первая пара (операнды, соединенные «&») – это собственно данные и способ кодирования, вторая – воспринятый текст и контекст, что в целом обеспечивает адекватность и восприятия, и понимания. Таким образом, можно сделать вывод, что иерархии форм материи должна соответствовать иерархия метаструктур, обеспечивающая согласованное использование, в том числе и межуровневое.

5. Переработка информации связана с соотнесением информации, воспринимаемых различий с объектами, которые передают эти различия в форме сигналов. «Такого соотнесения, по-видимому, нет в неживой природе, отражение носит там пассивный характер (хотя на каком-то этапе эволюции, вероятнее всего, химической, оно обретает некоторую активность в ходе самоорганизации)» [39]. При этом целесообразность, обычно являющаяся изначальной составляющей в кибернетических системах, имеет своим следствием оценку реакции приемника на по-

[30] В общем случае приемником может быть и источник, например, в другой момент времени.

лученную информацию. Именно это соотнесение с контекстом (целью, обстоятельствами и т.п.) и порождает специальное свойство – смысл, а с учетом значимости соответствующего контекста появляется возможность ввести и измерить ценность информации.

6. Информация (как категория наравне с материей и энергией) – это основа организации, ее средство и среда: «Информация пронизывает все уровни организации материи и энергии. … Она является первопричиной движения материи и энергии и определяет направление этого развития в пространстве и времени» [40], т.е. информация – это не только то, что является отображением, но и то, что организует изменения в мире материальных и абстрактных сущностей.

СПИСОК ЛИТЕРАТУРЫ

1. Маркс К., Энгельс Ф. Сочинения. Том 23. – М.: Государственное изд-во политической лит-ры, 1960. – 908 с.
2. Пенроуз Р. Путь к реальности или законы, управляющие Вселенной. – М: Ижевск, 2007. – 912 с.
3. Бриллюэн Л. Научная неопределенность и информация. – М.: Книжный дом «ЛИБРОКОМ», 2019. – 272 с.
4. Берг А.И., Спиркин А.Г. Кибернетика и диалектико-материалистическая философия // Проблемы философии и методологии современного естествознания. – М.: Наука, 1973. – С. 139-146.
5. Винер Н. Кибернетика. – М.: Наука, 1983. – 340 с.
6. Бейтсон Г. Разум и природа: Неизбежное единство. – М.: Книжный дом «ЛИБРОКОМ», 2016. – 256 с.
7. Урсул А.Д. Отражение и информация. – М.: Мысль, 1973. – 231 с.
8. Кравец А.С. Природа вероятности. – М.: Мысль, 1976. – 173 с.
9. Эшби У.Р. Введение в кибернетику. – М.: КомКнига, 2006. – 432 с.
10. Урсул А.Д. Природа информации: философский очерк. – Челябинск: Челяб. гос. академия культуры и искусств, 2010. – 231 с.
11. Гуревич И.М. Законы информатики – основа строения и познания сложных систем. – М.: ТОРУС ПРЕСС, 2007. – 400с.
12. Глушков В.М. О кибернетике как науке // Кибернетика, мышление, жизнь. – М.: Мысль, 1964. – С. 53-54.
13. Гуревич И.М. О физической информатике: предпосылки и основные результаты. – М.: ЛЕНАНД, 2014. – 160 с.
14. Zeilinger A.A. A Foundational Principle for Quantum Mechanics // Foundations of Phisics. – 1999. – Vol. 29, № 4. – P. 631-643.
15. Розенталь И.Л., Архангельская И.В. Геометрия, динамика, Вселенная. – М.: УРСС, 2003. – 200 с.
16. Князева Н.Н., Курдюмов С.П. Основания синергетики. – М.: КомКнига, 2006. –232 с.
17. Пригожин И., Стенгерс И. Время, хаос, квант. – М.: Едиториал УРСС, 2003. –240 с.
18. Чернавский Д.С. Синергетика и информация (динамическая теория информации). – М.: Едиториал УРСС, 2004. – 288 с.
19. Einstein A., Podolsky B., Rosen N. Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? // Phys. Rev. – 1935. – Vol. 47, Iss. 10. – P. 777-780.
20. Татур В.Ю. Тайны нового мышления. – М.: Изд-во «Прогресс», 1990. – 199 с.
21. Zurek W.H. Quantum Darwinism // Nature Physics. – 2009. – Vol. 5. – P. 181–188.
22. Ландау Л.Д., Лифшиц Е.М. Теоретическая физика: учебное пособие для вузов. В 10 т. Т. III. Квантовая механика (нерелятивистская теория). 4-е изд., испр. – М.: Наука, 1989. – 768 с.
23. Девитт Б. Квантовая механика в интерпретации Эверетта // Наука и предельная реальность: квантовая теория, космология и сложность. – М.: Ижевск, 2013 .— С.143-168.
24. ГОСТ Р ИСО/МЭК 10746-2-2000. Информационная технология. Взаимосвязь открытых систем. Управление данными и открытая распределенная обработка. Часть 2. Базовая модель. – М.: Стандартинформ, 2006. – 23 с.
25. ISO/IEC 2382:2015. Information technology — Vocabulary. – URL: https://www.iso.org/obp/ui/#iso:std:iso-iec: 2382:ed-1:v1:en (дата обращения 08.06.2020).
26. ГОСТ 7.0-99. Информационно-библиотечная деятельность, библиография. Термины и определения. – М: ИПК Изд-во стандартов, 1999. – 24 с.
27. ISO/IEC/IEEE 24765:2010. Systems and software engineering — Vocabulary. – URL: http://www.cse.msu.edu/~cse435/Handouts/Standards/IEEE24765.pdf (дата обращения 08.06.2020).
28. Хакен Г. Информация и самоорганизация. Макроскопический подход к сложным системам. – М.: КомКнига, 2005. – 248 с.
29. Pierce Ch.S. The Collected Papers of Charles Sanders Pierce. – London: Thoemmes Continuum, 1998. – 352 p.
30. Эко У. Роль читателя. Исследования по семиотике текста. – М.: Из-во РГГУ, 2005. – 502 с.
31. Спиноза Б. Этика. Ч.1. О Боге. Аксиомы. – М.: Фолио, 2001. – 388 с.
32. Информационный подход в междисциплинарной перспективе (круглый стол) // Вопросы философии. – 2010. – № 2. – С. 84-112.
33. Голицына О.Л., Максимов Н.В., Попов И.И. Информационные системы : учеб. пособие. – М.: Форум, 2007. – 496 с.
34. Гладких Н.Г. Динамические информационные процессы // Системы и средства информатики. – 2011. – Вып.11. – С. 341-362.
35. Винер Н. Я математик. – М.: Наука, 1967. – 356 с.
36. Максимов Н.В. Информация и знания: природа, концептуальная модель // Научно-техническая информация. Сер. 2. – 2010. – №7. – С. 1-10.

37. Максимов Н.В., Голицына О.Л., Монанков К.В., Лебедев А.А., Баль Н.А., Кюрчева С.Г. Средства семантического поиска, основанные на онтологических представлениях документальной информации // Научно-техническая информация. Сер. 2. – 2019. – №7. – С. 8-19.
38. Стратонович Р.Л. Теория информации. – М. Сов. радио, 1975. – 424 с.
39. Гуревич И.М., Урсул А.Д. Информация – всеобщее свойство материи. Изд. 2-е. – М.: КД «ЛИБРОКОМ», 2013. – 312 с.
40. Колин К.К. Сущность информации и философские основы информатики // Информационные технологии. – 2005. – № 5. – С. 63-70.

Материал поступил в редакцию 08.06.2020

Сведения об авторах

МАКСИМОВ Николай Вениаминович – доктор технических наук, профессор, профессор кафедры финансового мониторинга Национального исследовательского ядерного университета МИФИ (НИЯУ МИФИ), Москва
e–mail: nv-maks@yandex.ru

ЛЕБЕДЕВ Александр Анатольевич – ведущий математик кафедры финансового мониторинга Национального исследовательского ядерного университета МИФИ (НИЯУ МИФИ), Москва.
e-mail: lebedevalex@live.ru

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 002:004.81

А.С. Баканов, Н.Б. Баканова

Использование нечетких когнитивных карт при проектировании информационных систем организационного управления*

Рассмотрена возможность использования нечетких когнитивных карт, для мониторинга и визуализации представлений всех участников команды разработчиков о функциях информационной системы при ее проектировании.

Ключевые слова: *информационные системы, нечеткие когнитивные карты, моделирование, проектирование информационных систем*

DOI: 10.36535/0548-0027-2020-07-2

ВВЕДЕНИЕ

В настоящее время информационные системы являются эффективным инструментом накопления и обработки данных, охватывающим различные стороны управленческой деятельности, включая документооборот, управление ресурсами, планирование, оперативное управление, мониторинг, контроль.

Информационные системы (ИС) обеспечивают поддержку внутренних функциональных задач управленческой организации, включают режимы электронного взаимодействия с вышестоящими и подведомственными организациями, предоставляют гражданам возможность электронного общения с организационными структурами (например, портал ССТУ – сетевой справочный телефонный узел создан для обработки обращений граждан), предусматривают взаимодействие с внешними информационными системами (включая МЭДО – межведомственный электронный документооборот, СМЭВ – система межведомственного электронного взаимодействия, межведомственный портал «МВ-портал» и др.)[1].

Развитие информационных технологий, накопившийся опыт работы сотрудников управленческих организаций с информационными системами, создаваемые в процессе работ информационные ресурсы выдвигают новые требования к расширению функциональности информационных систем. В системы включаются программные средства, реализующие инновационные направления, связанные с созданием объединенных информационных хранилищ данных, с анализом и прогнозированием отраслевого развития, с использованием режимов поддержки принятия решений.

Таким образом, современные информационные системы поддержки организационного управления становятся достаточно сложными программно-техническими комплексами, которые используются специалистами различных функциональных направлений. В связи с этим вопросы, связанные с созданием, проектированием и поддержкой информационных систем, неизбежно усложняются [1, 2].

Основным инструментом разработки любых сложных систем, в том числе и программных, является моделирование. На всех этапах жизненного цикла ИС необходимо иметь её визуализированную актуальную структуру, на которой полно и всесторонне может быть отражено функционирование ИС, представлены связи между функциональными модулями, показано взаимодействие системы с внешним миром. Разработка методов и методик, позволяющих адекватно представлять и визуализировать структуру крупномасштабных информационных систем, – это актуальная научная задача [3, 4].

Настоящая статья посвящена одному из аспектов проектирования ИС, а именно расширению инструментальных средств моделирования визуального представ-

*Исследование выполнено в рамках Государственного задания Минобрнауки РФ по темам: № 0017-2019-0005 «Теоретические и прикладные проблемы информационных технологий, реалистичной компьютерной графики, визуальной аналитики и обработки многомерных данных», № 0159-2020-0001 «Психологические проблемы профессионального менталитета в условиях организационных и технологических инноваций».

[1] Федеральная государственная программа "Информационное общество (2011 – 2020 годы)", утверждена распоряжением Правительства РФ № 1815-р от 16.11.2010 г.

Указ Президента Российской Федерации от 17.04.2017 г. № 171 «О мониторинге и анализе результатов рассмотрения обращений граждан и организаций»

ления системы за счет использования когнитивных карт для выявления и интеграции функциональных задач управленческой организации в рамках этой системы.

МЕТОДЫ ПРОЕКТИРОВАНИЯ ИНФОРМАЦИОННЫХ СИСТЕМ

Эволюция информационных систем, усложнение и развитие их функциональных возможностей определяют сложность используемых программных комплексов, и, в свою очередь, формируют высокие требования к процессам проектирования систем. Все методы проектирования крупномасштабных распределенных информационных систем базируются на использовании моделей. Для быстрого и эффективного представления и восприятия информации в моделях следует использовать графические изображения. Большинство из существующих методологий проектирования информационных систем обеспечивает возможность представления объектов моделирования в виде различных графических нотаций для визуализации модели.

Выбор конкретной методологии при разработке модели определяется взаимосвязанной совокупностью ряда факторов, включающих навыки и опыт разработчиков. В процессе создания методологий проектирования был сделан вывод о целесообразности разработки графических языков для моделирования информационных систем, поскольку описательные языки не обладают достаточной наглядностью и эффективностью восприятия. В качестве обоснования этого вывода указывалось, что описательные языки не обеспечивают достаточного уровня непротиворечивости и полноты описания, в то время, как это является одним из обязательных требований в процессе разработки и проектирования ИС. Применение различных типов визуализации, графических нотаций, использующихся при проектировании, обусловлено большим количеством классов задач, на решение которых ориентированы те или иные модели.

В настоящее время широко используемыми подходами являются функционально-модульный и объектно-ориентированный [5, 6]. Функционально-модульный подход основан на использовании системного структурного анализа, при котором исследуемая система рассматривается как совокупность взаимосвязанных блоков или элементов, работа которых направлена на достижение общей цели. Появившаяся в середине 60-х гг. XX в. технология моделирования *SADT* (*Structured Analysis and Design Technique*) использовала функционально-модульный подход. В частности, технология моделирования *SADT* применялась Военно-воздушными силами США в программе интеграции компьютерных и промышленных технологий (*Integrated Computer Aided Manufacturing, ICAM*) и получила обозначение *IDEF*0 (*Icam DEFinition*) [6].

В качестве нового способа для проектирования в *SADT* был предложен графический язык для описания информационных и бизнес-процессов, представляющий совокупность упорядоченных, структурированных и взаимосвязанных графических диаграмм.

В ходе исследований методов проектирования разработан целый ряд графических нотаций для визуализации представления информационных систем. Наиболее известными являются:

- *IDEF*0 – для документирования процессов прохождения информации, использования ресурсов на каждом из этапов проектирования систем;
- *IDEF*1 – для документирования информации о производственном окружении систем;
- *IDEF*2 – для документирования поведения системы во времени;
- *IDEF*3 – для моделирования бизнес – процессов;
- *DFD* – диаграмма потоков данных (*data flow diagram*) – один из основных инструментов, описывающих внешние по отношению к системе источники и адресаты данных, логические функции, потоки данных и их хранилища, к которым осуществляется доступ.

Графические нотации показали свою эффективность для моделирования информационных процессов. На их основе реализованы первые инструментальные средства, предназначенные для автоматизации проектирования информационных систем на основе технологии моделирования *SADT*.

Отличие между функционально-модульным и объектно-ориентированным подходами состоит в различных способах декомпозиции. Объектно-ориентированный подход использует объектную декомпозицию, т.е. структура информационной системы описывается в терминах объектов и связей между ними. Функционирование ИС представляется посредством информационного обмена между объектами. При функционально-модульном подходе используется функциональная декомпозиция – функционирование системы описывается в терминах иерархии функций (передачи информации между отдельными функциональными модулями ИС). При объектном подходе, который основан на принципах абстрагирования, модульности и полиморфизма, объектная модель представляет объект моделирования в виде понятий «сущность-связь».

К достоинствам функционально-модульного подхода следует отнести возможность проектирования ИС «сверху вниз», что соответствует традиционным представлениям об иерархии функций информационной системы и понятно специалистам прикладной области. Недостаток функционально-модульного подхода заключается в достаточно сложном и неоднозначном переходе от иерархии функций ИС к разработке структуры программного обеспечения, что приводит к необходимости использования других методов представления системы и, соответственно, программных средств автоматизации проектирования. Применение различных видов диаграмм и переход к другим программным средствам усложняет работу проектировщиков и увеличивает время разработки и трудоемкость создания ИС [7].

При использовании функционально-модульного подхода важен правильный выбор объектов системы и их дальнейшая детализация. Критериями для такого выбора являются принципы оптимизации и повторного использования модулей построения программного комплекса ИС в целом. Эти критерии и

принципы не всегда очевидны для специалистов предметной области [8].

Для проектировщиков ИС, в свою очередь, существенные трудности представляет прикладная область, поскольку она имеет свою специфику и ряд особенностей, которые подчас не формализованы в техническом задании. Поэтому эффективная коллективная деятельность специалистов прикладной области и проектировщиков ИС при разработке крупномасштабных информационных систем является залогом успешной реализации проекта. Функционально-модульный подход более прост для понимания и использования как для специалистов прикладной области, так и для ИТ-специалистов, что позволяет проводить совместное проектирование и анализ функционирования ИС.

МОДЕЛИ, ИСПОЛЬЗУЕМЫЕ НА ЭТАПАХ ИССЛЕДОВАНИЯ ПРЕДМЕТНОЙ ОБЛАСТИ

При создании системы автоматизации модель предметной области является упрощенным представлением задач и функций организации. Для реализации процессов проектирования такое представление может быть рассмотрено на следующих моделях системы:

- логическое представление информационной системы, включающее, например, структурное представление (объекты, функции, задачи, информационные потоки), описательное представление и т.д.;
- функциональное представление информационной системы, предназначенное для исследования и анализа ее динамических характеристик или отдельных функциональных элементов (в динамике).

На логическом уровне предметную область возможно представить в виде следующих моделей:

- структурной модели, предназначенной для выявления элементов объекта информатизации, существенных с точки зрения поставленных задач и определения их взаимосвязей;
- информационной модели, применяемой для исследования внутренних и внешних информационных потоков;
- функциональной модели, предназначенной для разработки функций и задач элементов объекта.

Динамические модели описывают функционирование объекта и предназначены для визуализации предметной области с учетом динамических характеристик информационных и технологических процессов организации. Изучение технологических процессов и задач организационного управления с помощью динамических моделей позволяет количественно оценить характеристики информационных потоков и распределение нагрузки на функциональные элементы организационной структуры, выявить «узкие места», возникающие при выполнении технологических операций, уточнить характеристики создаваемой системы, оказывающие существенное влияние на эффективность выполнения работ [10]. Краткий перечень динамических характеристик информационных процессов, которые необходимо исследовать при создании крупномасштабной распределенной ИС, представлен в табл. 1.

Понятие «жизненный цикл системы» определяет основные этапы эволюции новой системы от замысла и концептуальной разработки до выведения из эксплуатации. Стадии жизненного цикла систем и жизненного цикла программных средств определены в стандартах ГОСТ Р ИСО МЭК 15288-2005 и ГОСТ Р ИСО/МЭК 12207-2010, согласно которым процессы жизненного цикла определяются, настраиваются и используются на всех его стадиях до полного достижения целей и результатов. В соответствии со стандартом единой универсальной модели жизненных циклов систем не существует. Те или иные стадии жизненного цикла могут отсутствовать или присутствовать в зависимости от каждого конкретного случая разработки системы.

Таблица 1

Перечень основных информационных процессов

№ п/п	Название исследуемой характеристики ИС	Единицы измерения
1	Внешние информационные потоки	Бит/сек
2	Виды и характеристики взаимодействия ИС с внешним миром	Количество, нотации описания
3	Внутренние информационные потоки	Бит/сек
4	Производительность оборудования, используемого в процессах обмена данными	Бит/сек

В процессе жизненного цикла с системой работают специалисты разного профиля: системные аналитики, проектировщики, разработчики, программисты различных направлений, специалисты поддержки функционирования (ИТ-специалисты), функциональные пользователи системы. Для каждой стадии жизненного цикла используются те принципы моделирования и те модели, которые позволяют наиболее полно отразить проблемы, возникающие на соответствующем этапе. На рис. 1 показаны виды моделей, которые могут использоваться в процессе жизненного цикла крупномасштабных систем.

На представленной схеме показана связь используемых видов моделей с этапами жизненного цикла информационных систем. Названия этих этапов даны согласно ГОСТ Р ИСО МЭК 15288-2005.

Этап жизненного цикла ***«замысел, обследование»*** предназначен для подготовки технического задания на проектирование системы. На этом этапе можно рассмотреть следующие модели: логическое представление объекта автоматизации и модели процессов функционирования.

Логическое представление объекта автоматизации применяется для визуализации и спецификации структурного построения системы. Оно формируется в процессе обследования и предназначается для анализа информационных потоков, логической взаимосвязи модулей, формализации функциональных задач, для подготовки технического задания. При использовании функционально–модульного подхода к представлению объекта последний описывается в

нотациях *AS-IS* (как есть) и *TO-BE* (как надо). Представление *AS-IS* описывает актуальное состояние информационных потоков и технологических процессов объекта. Представление *TO-BE*, как правило, предлагает различные варианты изменения потоков и процессов, которые должны быть предусмотрены техническим заданием на создание или модернизацию ИС. В дальнейшем при исследовании и сравнении разработанных моделей можно выявить узкие места информационных потоков, непроизводительные затраты, неэффективные технологические процессы. Анализ моделей *AS-IS* и *TO-BE* позволяет предложить оптимизационные схемы информационных потоков, выполнения работ, повысить производительность информационной системы [6, 8].

Модели процессов функционирования являются, как правило, динамическими моделями и предназначены для расчета динамических характеристик отдельных блоков или режимов системы, которые также должны быть учтены в техническом задании. Так, например, для крупных информационных систем на моделях могут быть проверены предполагаемые потоки данных, число рабочих мест обработки данных, пропускные характеристики линий связи, характеристики аппаратуры приема/передачи данных и другие.

Этап ***«проектирование, создание»*** предназначен непосредственно для проектирования системы. На этом этапе используются, как правило, средства автоматизации проектирования, так называемые *CASE*-средства (*Computer Aided Software Engineering*), в которых реализованы функции поддержки коллективной разработки систем, имеются встроенные средства документирования, а также ряд других сервисных возможностей.

Программные *CASE*-средства, реализующие объектно-ориентированный подход, позволяют использовать модели «потоков данных», «потоков работ» и т.п. В них реализованы средства согласования, позволяющие осуществлять переход от одного процесса к другому, графически представлены последовательные и параллельные процессы. Имеются также средства для моделирования размещения программного комплекса. В качестве примера можно привести язык проектирования *UML* (*Unified Modeling Language*), использующий более 20 типов визуальных диаграмм. Каждая диаграмма представляет некоторый взгляд на создаваемую систему. Язык *UML* предназначен для визуализации информационных процессов и может быть использован как ИТ-специалистами, проектировщиками ИС, так и программистами прикладной области [7].

На следующем этапе жизненного цикла ***«применение и сопровождение»*** прикладная область задачи рассматривается совместно с используемой информационной системой, поэтому возникают новые представления прикладной области, составной частью которой являются автоматизированные процессы. Модели этого этапа предназначены, как правило, для предложений по развитию системы. Предложения по совершенствованию ее отдельных функций возникают у специалистов прикладной области в процессе изменения структуры организации, освоения возможностей ИС, накопления информационной базы, появления новых технических средств приема/передачи информации и т.п. Предложения по развитию ограничиваются конкретными прикладными задачами, которые считаются приоритетными, но при этом не касаются программного комплекса в целом.

В табл. 2 показана целесообразность использования различных видов моделей при создании крупномасштабных распределенных информационных систем.

Рис. 1. Виды моделей, используемых при создании информационных систем

Таблица 2

Этапы разработки информационных систем

№	Наименование этапа	Статика	Динамика
1	Обследование предметной области и формирование технического задания	v	
2	Концептуализация предметной области и разработка концептуального представления информационной системы	v	
3	Разработка и изучение основных функциональных режимов на имитационных или математических моделях		v
4	Корректировка концептуального представления информационной системы по результатам моделирования	v	v
5	Разработка прототипов системы и моделирование функциональных режимов на прототипах системы	v	v
7	Сопровождение жизненного цикла системы	v	v
8	Сбор и подготовка данных для расширения функционала и модернизации системы.	v	v

Примечание: В колонках «статика» и «динамика» отмечены типы моделей, позволяющие наиболее полно отразить особенности работы системы на соответствующих этапах жизненного цикла. Статические модели реализуются в виде визуальных диаграмм с использованием соответствующих программных средств и методологии и, по сути, представляют документированный проект информационной системы, представленный с использованием той или иной методологии.

Из табл. 2 видно, что статические модели рекомендуется создавать на всех этапах разработки информационной системы. Основное предназначение статических моделей – описывать и визуализировать структуру и архитектуру ИС. Под статическими моделями иногда понимают совокупность статических диаграмм, представленных в функционально-модульной или объектной парадигме; такие модели применяются для визуализации процесса разработки программного комплекса ИС. Однако они не предназначены для анализа и представления динамических процессов, не позволяют моделировать потоки данных и другие динамические характеристики системы [8].

Динамические модели предназначены для анализа поведения системы. Их рекомендуется использовать в процессах разработки – на этапах 3 и 4 (см. табл. 2). Обычно динамические модели создаются уже после статических моделей, описывающих основные представления системы.

Результаты моделирования, полученные при исследовании статических и динамических моделей, позволяют внести существенные коррективы в проект системы. Динамические модели могут дорабатываться и уточняться в процессе жизненного цикла ИС на этапах сопровождения и модернизации системы.

При использовании объектно-ориентированного подхода к проектированию информационных систем можно использовать средства, предназначенные для моделирования взаимодействия программных модулей. Однако разработанные с использованием таких средств модели свидетельствуют только о корректности понимания разработчиками назначения системы и не способны выявить ошибки, связанные с разработкой проекта системы и постановкой задачи. Устранение ошибок, возникающих при разработке программных модулей, происходит, как правило, на этапах отладки и опытной эксплуатации. Ошибки, допущенные при разработке программного кода, оказывают в дальнейшем значительно меньшее влияние на корректную работу самой информационной системы, чем ошибки, допущенные на этапе разработки проекта системы. Своевременно не исправленные ошибки, допущенные на этапе постановки задачи, устраняются значительно сложнее и зачастую могут привести к отрицательному результату или невозможности внедрения и использования ИС [6, 8].

В рассмотренных подходах к проектированию информационных систем (функционально-модульный и объектно-ориентированный) достаточно хорошо проработаны возможности отражения основных требований к функциональности системы. При этом надо учитывать, что все требования на первоначальных этапах разработки имеют обобщенный, нечеткий, иногда противоречивый, слабоструктурированный характер. Наличие недетерминированных требований зачастую является основной причиной появления «неуспешных проектов» или существенных проблем, возникающих в процессе реализации проекта ИС [6, 8].

Для минимизации рисков необходимо проводить обследование информационной системы с помощью моделей функционирования, разработка которых в процессе проектирования достаточно трудоемка и поэтому оправдана только для крупномасштабных информационных проектов. По этой причине такие модели не включены в методологические подходы проектирования информационных систем.

Однако разработка моделей функционирования, адекватно отражающих информационные и технологические процессы, оправдана на этапах проектирования крупномасштабных ИС со сложной архитекту-

рой. Модели используются, как правило, в процессе обследования предметной области для проверки концептуальных решений проекта. Такие модели позволяют количественно оценить входные и выходные потоки данных, изучить структуру, особенности поведения и выявить специфику объекта информатизации.

Независимо от выбранной методологии проектирования и средств визуализации создание информационной системы начинается с обследования предметной области, в результате чего создается ее концептуальная модель в статическом виде. На последующих этапах разрабатываются модели программной реализации, которые могут быть представлены как в статическом, так и в динамическом виде. Использование только статических или только динамических моделей может негативно отразиться на создаваемом проекте в целом, что в дальнейшем не позволит его успешно реализовать.

На начальном этапе проектирования крупной многофункциональной ИС достаточно сложно определить детальные требования к функциональным задачам в рамках общей задачи информатизации организационной структуры, поскольку для конечных пользователей недостаточно ясны возможности системы. Любые требования и пожелания необходимо сопоставлять с ее общим функциональным представлением, включая многозадачность, учитывая при этом работу специалистов других прикладных областей, системные ограничения и ограничения по выполнению задач сопровождения. Поэтому вначале определяются и автоматизируются основные требования к обработке информационных потоков, накоплению данных, созданию информационных хранилищ.

КОГНИТИВНЫЕ КАРТЫ ДЛЯ ВИЗУАЛИЗАЦИИ ФУНКЦИЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Расширение спектра задач, решаемых в управленческих организациях, развитие функционала информационных систем, активное использование процедур взаимодействия с вышестоящими и подведомственными организациями, внедрение режимов поддержки принятия решений – все это приводит к трансформации информационных систем организационного управления в крупномасштабные распределенные информационные комплексы.

Интеграция прикладных задач в рамках информационного комплекса организационной структуры (единый информационный пул), с одной стороны, повышает непротиворечивость данных, но, в то же время, повышает сложность системы, как с точки зрения программного комплекса, так и с точки зрения представлений пользователями функционала системы. Дополнительным фактором, усложняющим работу системы, является рост количества пользователей, подключение дополнительных устройств, добавление новых режимов взаимодействия, сервисов обработки данных, поддерживаемых современными информационными технологиями. Все это определяет необходимость постоянного анализа и мониторинга функционирования систем для поддержки их в актуальном состоянии и для своевременной подготовки проектов развития.

Апробированный и хорошо зарекомендовавший себя подход для исследования предметной области и разработки информационных систем – использование разнообразных моделей, которые поддерживаются методами, описанными в предыдущих разделах настоящей статьи.

Дополнительным эффективным инструментом для мониторинга и моделирования процессов, протекающих в крупномасштабных распределенных комплексах, являются когнитивные карты и нечеткие когнитивные карты [7]. Они позволяют отразить взгляды различных пользователей на систему с точки зрения ее функционирования, так как один из отличительных элементов, характеризующий крупномасштабную информационную систему, – распределенность. Пользователи могут находиться в разных территориально удаленных точках и взаимодействовать между собой посредством информационной системы. На процесс работы влияют не только ее характеристики, но также и представления о них удаленных пользователей. Поэтому для подготовки проекта необходимо проанализировать взгляды различных пользователей на функционал системы и задачи ее развития. В этом случае эффективным инструментом являются когнитивные карты, которые представляют графовую модель сложной ситуации в виде причинно-следственных связей между факторами, основанную на экспертных представлениях [9–12].

В общем случае когнитивные карты (от лат. *cognitio* — знание, познание) предназначены для отражения субъективных представлений исследователя или группы исследователей о пространственной организации внешнего мира. Они создаются и видоизменяются в результате активного взаимодействия субъекта с окружающим миром. Когнитивные карты первоначально были предложены Р. Аксельродом в 1976 г. для визуализации представления «сущность-связь» [13]. Позднее Б. Коско выдвинул идею использовать в когнитивных картах элементы нечёткой логики (англ. *fuzzy logic*) [14].

В настоящее время нечеткие когнитивные карты предоставляют широкие возможности для мониторинга и моделирования сложных, крупномасштабных распределенных систем. В проведенном нами исследовании для изучения представлений различных ***групп пользователей*** о некоторой распределенной информационной системе, ее взаимосвязях и функциях использовался описанный ниже подход [15, 16].

Взгляд пользователя на процесс функционирования системы отражается в виде нечёткой когнитивной карты, представленной графом, вершинами которого являются модули (объекты) распределенной системы (функции, сервисы и возможности, предоставляемые этими модулями), а дуги графа – это функциональные связи между объектами когнитивной карты. На карте указываются связи следующих видов: явная связь, неявная (возможная – по субъективному мнению пользователя) связь. Веса дуг варьируются в диапазоне от 0 до 1 [17].

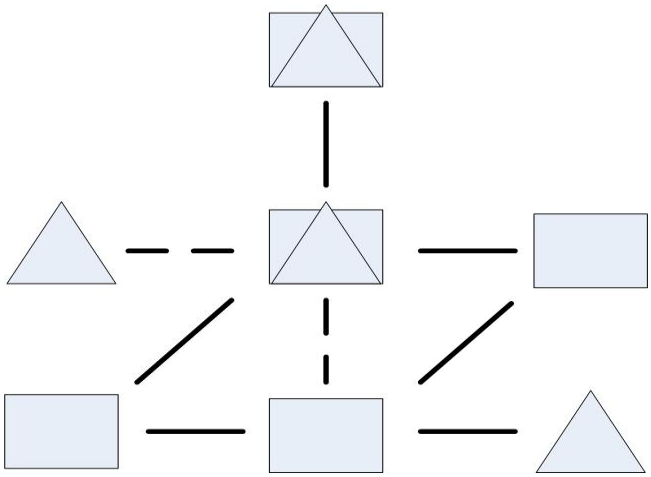

Рис. 2. Нечеткая когнитивная карта, визуализирующая представления различных групп пользователей об объекте автоматизации

Нечеткая когнитивная карта, визуализирующая представления различных ***групп пользователей системы*** о распределенной ИС представлена на рис. 2. ИТ-специалисты изображали объекты в виде прямоугольников, в то время как специалисты прикладной области – в виде треугольников. Таким образом, нечеткая когнитивная карта, представленная на рис. 2, является, по сути, композицией из двух когнитивных карт, одна из которых визуализирует представления ИТ-специалистов, а другая – представления специалистов некоторой прикладной области.

На рис. 2 сплошными линиями представлены связи между объектами и функциями системы. Пунктирной линией изображена возможная (согласно субъективному мнению) связь между модулями или сервисами распределенной системы.

Когнитивные карты визуализируют ментальную репрезентацию индивидуума о распределенной системе. Необходимо отметить, что существует достаточно большое количество определений термина «ментальная репрезентация». В нашей статье используется следующее определение: ментальная репрезентация – субъективный образ объективной реальности, отражение внутреннего и внешнего мира в сознании человека [3, 18] или, применительно к настоящему исследованию, субъективный структурированный образ об информационной системе, представляемый пользователем.

В исследованиях О.И. Ларичева и А.Б. Петровского отмечается, что в ходе взаимодействия с информационной средой специалисту приходится учитывать большое число различных факторов, а также решать задачи многокритериального выбора. Это приводит к нагрузке на систему переработки информации, вынуждая индивида использовать разные, порой весьма оригинальные эвристики для решения поставленных задач [19, 20]. Возможности человека по приему и переработке информации с позиций когнитивной психологии описываются с помощью различных функциональных моделей структуры памяти пользователя, механизмов процесса мышления и других познавательных процессов.

Использование когнитивных карт для описания представлений различных пользователей о функциях и задачах некоторой системы целесообразно для консолидации этих представлений с целью разработки функций системы. Применение такого подхода позволило создать модель деятельности ***группы прикладных пользователей*** некоторой распределенной крупномасштабной информационной системы. В задачу такой группы входило: разработать технические задания на модернизацию отдельных функциональных модулей информационной системы; моделировать взаимодействие между ее различными модулями; осуществить мониторинг функционирования разработанных модулей. Для этого ***группе прикладных пользователей системы*** необходимо было иметь представление о функциях и возможностях как всей системы в целом, так и о функциях и возможностях ее отдельных модулей. После распределения заданий между различными модулями системы пользователю следовало осуществить мониторинг выполнения заданий, а также анализ и оценку эффективности выполнения.

В процессе проведения исследований был выявлен ментальный образ информационной системы у разных групп пользователей. Он имел иерархическую структуру, которую можно оценить количественно и качественно. Структура имела вид ориентированного или направленного графа, узлы которого соответствовали функциональным модулям, а направленные дуги или ребра – связям, которые выявлялись каждым пользователем в процессе работы с информационной системой.

Количество уровней иерархии в структуре и количество дуг, сходившихся к одному узлу, характеризовали как структуру ментальной репрезентации участника группы, так и характер (особенности) его деятельности, опосредованный информационной системой. Например, если большое количество дуг сходились к одному узлу, то очевидно, что испытуемый пользователь преимущественно использует функционал именно этого модуля информационной системы.

Таким образом, визуализируя свое восприятие информационной системы, испытуемый визуализировал в виде графа и свою деятельность, опосредованную этой системой.

ЗАКЛЮЧЕНИЕ

При разработке моделей информационных систем возможно применять разнообразные графические нотации, отражающие логическое построение, структуру, функционирование ИС. Структурированная совокупность графических нотаций представляет определенную методологию и обеспечивает визуализацию создаваемого проекта.

Программные средства, на основе которых реализованы соответствующие методологии, ускоряют процесс разработки и позволяют быстро визуализировать и исследовать как статические, так и динамические представления предметной области объекта проектирования. Самостоятельная разработка динамических моделей в каждом конкретном случае часто нецелесообразна ввиду значительных временных и трудовых затрат. Поэтому они разрабатываются для конкретных классов крупномасштабных распределенных систем. В таких моделях выделяют и деталь-

но описывают специфичные характеристики объекта или некоторой его составной части. Использование динамических моделей позволяет исследовать определяющие характеристики, которые влияют на эффективность функционирования и структуру проектируемой ИС системы.

Однако длительный опыт показывает, что нельзя ограничиваться созданием только одной модели. Наилучший подход при разработке любой системы – использовать совокупность нескольких моделей.

В настоящей статье была рассмотрена возможность использования когнитивных карт и нечетких когнитивных карт для мониторинга и моделирования процессов, протекающих в крупномасштабных распределенных системах. Когнитивные карты являются эффективным инструментом, позволяющим отразить взгляды различных пользователей на систему с точки зрения ее функционирования.

СПИСОК ЛИТЕРАТУРЫ

1. Друкер П.Ф. Управление в обществе будущего. – М.: Вильямс, 2007. – 320 с.
2. Гиляревский Р.С., Залаев Г.З., Родионов И.И., Цветкова В.А. Современная информатика: наука, технология, деятельность. – М.: ВИНИТИ, 1998. – 220 с.
3. Журавлев А.Л. Психология управленческого взаимодействия. – М.: Институт психологии РАН, 2004. – 475 с.
4. Занковский А.Н. Психология лидерства: от поведенческой модели к культурно-ценностной парадигме – М.: Институт психологии РАН, 2011. – 296 с.
5. Баканова Н.Б. Анализ информационных процессов в управленческих организациях для реализации режимов поддержки принятия решений // Электросвязь. – 2015. – № 5. – С. 59-62.
6. Йордан Э., Аргила К. Структурные модели в объектно–ориентированном анализе и проектировании – М.: ЛОРИ, 1999. – 264 с.
7. Вендров А.М. Проектирование программного обеспечения экономических информационных систем: учебник. – М.: Финансы и статистика, 2005. – 544 с.
8. Буч Г., Рамбо Д., Якобсон И. UML. Руководство пользователя. Унифицированный язык моделирования. – М.: ДМК, 2000. – 432 с.
9. Холодная М.А. Когнитивные стили: О природе индивидуального ума. 2-е изд. – СПб: Питер, 2004. – 384 с.
10. Баканов А.С., Зеленова М.Е., Алдашева А.А. Когнитивный стиль как фактор надежности работы в системе электронного документооборота // Социальные и гуманитарные науки на Дальнем Востоке. – 2015. – № 3(47). – С. 61-67.
11. Сиваш О.Н., Баканов А.С., Зеленова М.Е. Моделирование информационного взаимодействия в системах человек- компьютер // Вестник Костромского государственного университета. Серия: Педагогика. Психология. Социокинетика. – 2017. – Т. 23, № 3. – С. 90-95.
12. Толочек В.А. Стили деятельности: ресурсный подход. – М.: Институт психологии РАН, 2015. – 366 с.
13. Axelrod R. Structure of decision: the cognitive maps of political elites. – Princeton: Princeton University Press, 1976.
14. Kosko B. Fuzzy Engineering. – NJ: Prentice Hall, 1996. ISBN 0-13-124991-6.
15. Фестингер Л. Теория когнитивного диссонанса. – СПб: Ювента, 1999.
16. Махнач А.В., Дикая Л.Г. Мировоззренческая направленность как компонент жизнеспособности человека в социономических профессиях // Организационная психология и психология труда. – 2018. – Т. 3, №1. – С 62-91.
17. Bakanov, A.S.; Zelenova, M E. Cognitive styles as determinants of success in professional activity // Social Psychology And Society. – 2015. – Vol. 6, № 2. – P. 61-75
18. Баканов А.С., Ташев Т.Д., Баканова Н.Б. Мониторинг крупномасштабной информационной системы с использованием нечетких когнитивных карт // Управление развитием крупномасштабных систем (MLSD'2019). Материалы двенадцатой международной конференции / под ред. С.Н. Васильева, А.Д. Цвиркуна. – М.: Международный научно-исследовательский институт проблем управления РАН, 2019. – С. 1039-1043.
19. Петровский А.Б. Теория принятия решений. – М.: Академия, 2009. – 400 с.
20. Петровский А.Б. Многокритериальное принятие решений по противоречивым данным: подход теории мультимножеств // Информационные технологии и вычислительные системы. – 2004. – № 2. – С. 56-66.

Материал поступил в редакцию 11.05.20.

Сведения об авторах

БАКАНОВ Арсений Сергеевич – кандидат технических наук, научный сотрудник Института психологии РАН, Москва
e-mail: arsb2000@pochta.ru

БАКАНОВА Нина Борисовна – доктор технических наук, доцент, ведущий научный сотрудник Института прикладной математики им. М.В. Келдыша РАН (ИПМ им. М.В. Келдыша РАН), Москва,
e-mail: nina@keldysh.ru

УДК 004.89:510.64

Д.К. Чебанов

Об особенностях реализации решателя ДСМ-метода для интеллектуального анализа данных*

Рассматривается программная реализация процедур ДСМ-метода автоматизированной поддержки исследований, ранее применявшегося для решения задач, связанных с прогнозированием заболеваний на основе различных данных, в том числе геномных.

Уделяется внимание приемам по оптимизации использования памяти и сокращению вычислительного времени, в том числе организации параллельного исполнения процедур. Разработка велась на языке python 3.7. Предложенная оптимизация позволит сократить время вычислительных процедур более чем в 20 раз.

Ключевые слова: *ДСМ-метод АПИ, организация вычислений, онкологические данные, искусственный интеллект, python, интеллектуальный анализ данных, оптимизация вычислений.*

DOI: 10.36535/0548-0027-2020-07-3

ВВЕДЕНИЕ

При помощи ДСМ-метода автоматизированной поддержки исследований (ДСМ-метод АПИ) [1] ранее мы решали задачи прогнозирования развития онкологического заболевания и определения новых патогенных мутаций по генетическим и медицинским данным. Успешно примененный ДСМ-метод АПИ реализует обнаружение закономерностей в сложноструктурированных эмпирических данных, содержащих причинно-следственные зависимости в неявном виде [2, 3]. Этот метод обеспечивает формализацию знаний предметной области средствами многозначной логики, для чего обобщает в гипотезах информацию, полученную из обучающей выборки, затем применяет эти гипотезы для предсказания исследуемого эффекта неизвестных объектов, а также имеет критерий достаточного основания правдоподобного вывода.

ДСМ-метод АПИ состоит из этапов: ДСМ-рассуждения, результатом которого являются гипотезы о причинах и предсказаниях; ДСМ-исследования, представляющего проверку полученных гипотез при последовательных расширениях исходной обучающей выборки (в терминах ДСМ-метода АПИ – база фактов).

ДСМ-метод АПИ реализуется при помощи интеллектуальной системы, которая состоит из базы фактов, базы знаний, пользовательского интерфейса и решателя задач – ДСМ-решателя. Последний включает рассуждатель, который реализует ДСМ-исследование. ДСМ-рассуждение осуществляется при помощи 16 стратегий, каждая из которых определена парой предикатов сходства: для (+)-примеров и для (–)-примеров (M^+-предиката и M^--предиката соответственно) [4].

Вычисленные при помощи ДСМ-метода комбинации признаков являются гипотезами о причинах исследуемого эффекта. Для минимизации случайности расширений базы фактов вычисления проводятся для всех возможных перестановок ее расширений. Если при этом гипотеза сохраняется, то она считается эмпирической закономерностью [5].

База фактов (БФ) представляла генотипические, фенотипические и иммунные данные 414 пациентов с диагнозом «меланома», из которых 216 имеют ремиссию заболевания на протяжении не менее двух лет, и считаются (+)-примерами, остальные 198 имеют рецидив и являются (–)-примерами. Задача состояла в предсказании состояния ремиссии пациентов по представленным данным, а также в выявлении таких комбинаций признаков, которые являются причинами данного состояния, либо его отсутствия.

В результате применения ДСМ-метода АПИ к онкологическим данным были достигнуты высокие метрики качества предсказания, а также получены полезные эмпирические закономерности, подтвержденные экспертами предметной области.

*Работа выполнена при финансовой поддержке РФФИ (проект № 18-29-03063)

В настоящей работе внимание уделяется программной реализации метода, в частности, организации процесса вычислений.

ВЫБОР СПОСОБА РЕАЛИЗАЦИИ

В качестве языка программирования для реализации ДСМ-метода АПИ был выбран *python* (версия 3.7). Этот язык относится к классу высокоуровневых объектно-ориентированных, а значит содержит большое количество встроенных возможностей, что на практике позволяет разработчику уделять значительно меньше внимания созданию базовых функций, сосредотачиваясь на реализации алгоритмов. В частности, программы, написанные на *python*, имеют сравнительно легко читаемый код с упрощенным синтаксисом и не требуют компилирования перед запуском, что упрощает работу программиста. Но, в то же время, необходимо больше вычислительных ресурсов, так как программы на *python* для своего выполнения требуют запуска ряда пакетов (вспомогательных программ).

В отличие от *python*, язык *C* является низкоуровневым, и в этой связи не требует выполнения промежуточных программ, «обращаясь» напрямую к процессору, что обуславливает выбор этого языка для написания большинства современных операционных систем и драйверов, критичных к быстродействию. У *python* имеется особенность, связанная со значительным по сравнению с *C* расходом памяти, – динамическая типизация переменных, при которой под объявленную переменную резервируется не строго необходимый объем памяти, а наибольший из возможных. Благодаря этим и другим особенностям программы, созданные на *C*, обладают более высокой производительностью по сравнению с созданными на других языках, в том числе на *python*. В подавляющем большинстве случаев именно *C* является выбором для высоконагруженным систем, действующих в условиях дефицита ресурсов.

Программная реализация ДСМ-метода АПИ предполагает поиск пересечений всех возможных сочетаний исходных примеров, что требует значительного количества ресурсов. Многие разработчики в этой связи выбирают для этой задачи языки семейства *C* [6]. Однако преимущество *python* заключается в его относительной простоте изучения, в том числе за счет минималистичного синтаксиса, и значительного количества подробно документированных библиотек, что и обеспечило ему популярность среди разработчиков. Этот язык изучается сегодня во многих учебных заведениях благодаря достаточной простоте его овладения. Так, веб-ресурс *PYPL*, занимающийся подсчетом рейтинга популярности языков программирования, поставил язык *python* на первое место в 2020 г.: его доля распространения почти в два раза превосходит второй по популярности язык *Java*.

Таким образом, смысл в реализации ДСМ-метода АПИ на языке *python* заключается в том, чтобы большее количество исследователей и разработчиков, в том числе не обладающих высокой степенью владения языками программирования, могли реализовать данный метод, применив его к собственным задачам.

ПАРАЛЛЕЛЬНОЕ ИСПОЛНЕНИЕ

Вычислительные процессы в представленной реализации ДСМ-метода АПИ организованы параллельно на трёх уровнях:

1) независимой друг от друга обработки (+)-примеров (пациентов с эффектом «ремиссия», имеющих истинностное значение «фактически истинно») и (–)-примеров (пациентов с эффектом «рецидив», имеющих истинностное значение «ложно»);

2) всех возможных перестановок расширений БФ: в каждом из потоков обрабатывается отдельная последовательность расширений БФ (16 потоков);

3) стратегий: каждая стратегия выполнялась в отдельном потоке (16 потоков).

Как было отмечено выше, обратной стороной высокоуровневой природы языка *python* является его более низкая скорость исполнения скриптов, в частности по сравнению с более низкоуровневым языком *C*. Тем не менее, при создании программ на языке *python* возможно организовать параллельное исполнение процессов, что может нивелировать вышеуказанный недостаток, особенно если учесть развитие вычислительной техники с многоядерной архитектурой процессора.

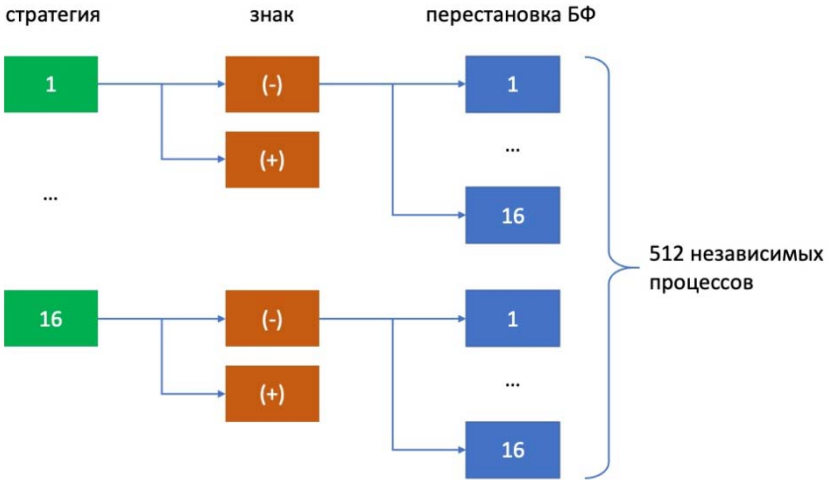

Рис. 1. Схема организации параллельных потоков на этапе порождения гипотез.

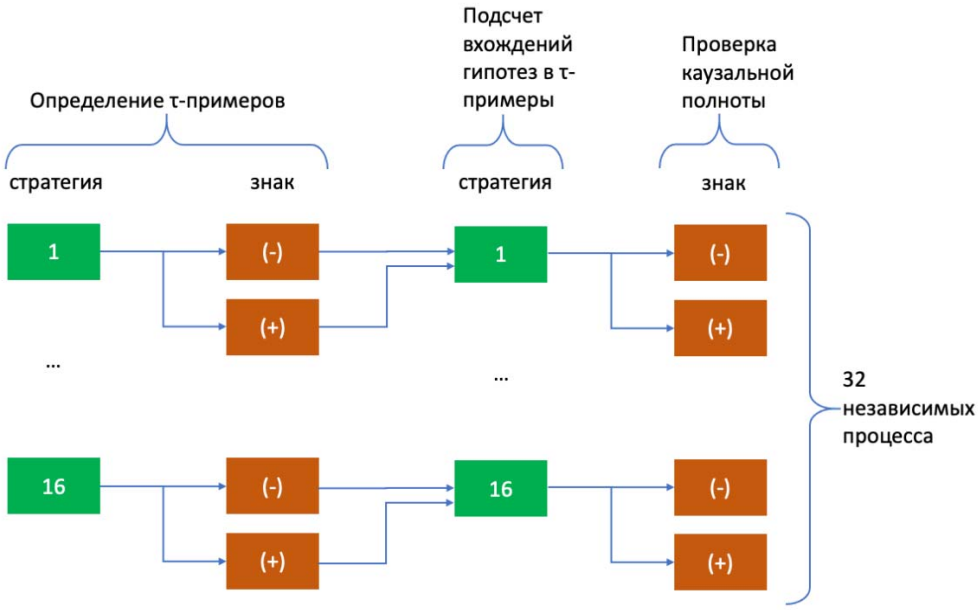

Рис. 2. Схема организации параллельных потоков на этапе предсказания примеров.

Для запуска параллельных процессов в языке *python* используется библиотека *multiprocessing,* в которой наиболее подходящий класс, предназначенный для выделения процесса в отдельный поток – *Pool.* Однако этот класс не поддерживает вложенные друг в друга уровни процессов, поэтому для реализации многоуровневой параллельной архитектуры программы был применен метод *starmap*, в котором в качестве аргументов использовались 512 наборов аргументов, соответствующих каждому из процессов. На рис. 1 представлена схема организации параллельных потоков исполнения программы на этапе порождения гипотез.

На этапе предсказания эффекта неизвестных примеров использовался другой подход: поскольку процесс несколько сложнее в плане организации вычислений и подразумевает изменение архитектуры при выполнении различных функций, для его реализации была применена комбинация классов *Pool* и *Process* библиотеки *multiprocessing* (рис. 2).

ОПТИМИЗАЦИЯ ПАМЯТИ ПРИ ВЫЧИСЛЕНИЯХ

Процесс определения пересечений примеров, как было указано ранее, является наиболее вычислительно сложной задачей в реализации ДСМ-метода АПИ. Ситуация с потреблением ресурсов памяти при реализации на языке *python* усугубляется динамической типизацией переменных, при которой разработчик не может управлять объемом оперативной памяти, выделяемой под хранение объектов вычислений.

В этой связи в качестве способа хранения примеров, каждый из которых представляет кортеж значений булевских переменных, был выбран массив типа *tuple*. В отличие от массива *list*, который предусматривает доступ к элементам по их индексу, массив *tuple* хранит только упорядоченный набор данных, не предполагающий изменения в нем порядка элементов, что позволяет ему занимать меньше памяти.

Реализация ДСМ-метода АПИ – это вычисление операции сходства на исследуемых примерах, что означает определение пересечения наборов признаков примеров. В программных реализациях ДСМ-метода АПИ применяется целый ряд алгоритмов, однако для задач представленного типа оптимальным с точки зрения ресурсов является алгоритм Норриса [7], он же используется исследователями в схожих задачах [6]. В представленной интеллектуальной системе реализован алгоритм Норриса с применением некоторых шагов по оптимизации работы с памятью. Он позволяет избегать вычисления тех комбинаций примеров, которые уже были вычислены ранее. Так, если ранее уже было вычислено пересечение $p(m \cap n)$ комбинации двух примеров m и n, то вместо операции пересечения $r(m \cap n \cap q)$ выполняется пересечение $r(p(m \cap n) \cap q)$, для которого значение $p(m \cap n)$ берется из памяти, что значительно экономичнее, особенно при комбинациях из большого количества примеров. Таким образом, для вычисления пересечения комбинации из m примеров используется массив, который хранит обработанные ранее комбинации из m-1 примеров (рис. 3 и *Листинг 1*). Для хранения полученных ранее гипотез используются массивы типа *set*, которые хранят неупорядоченные данные и являются наименее емкими с точки зрения занимаемой памяти.

Операция определения пересечений наиболее трудоемкая, к тому же она выполняется на представленной выборке порядка 10^{18} раз. В этой связи она должна быть максимально экономичной с точки зрения вычислительных ресурсов, поэтому пересечение примеров определяется при помощи приема обработки массивов *list comprehension*, при этом вместо объектов *list* используются объекты *tuple*, которые, как было отмечено, занимают меньший объем памяти.

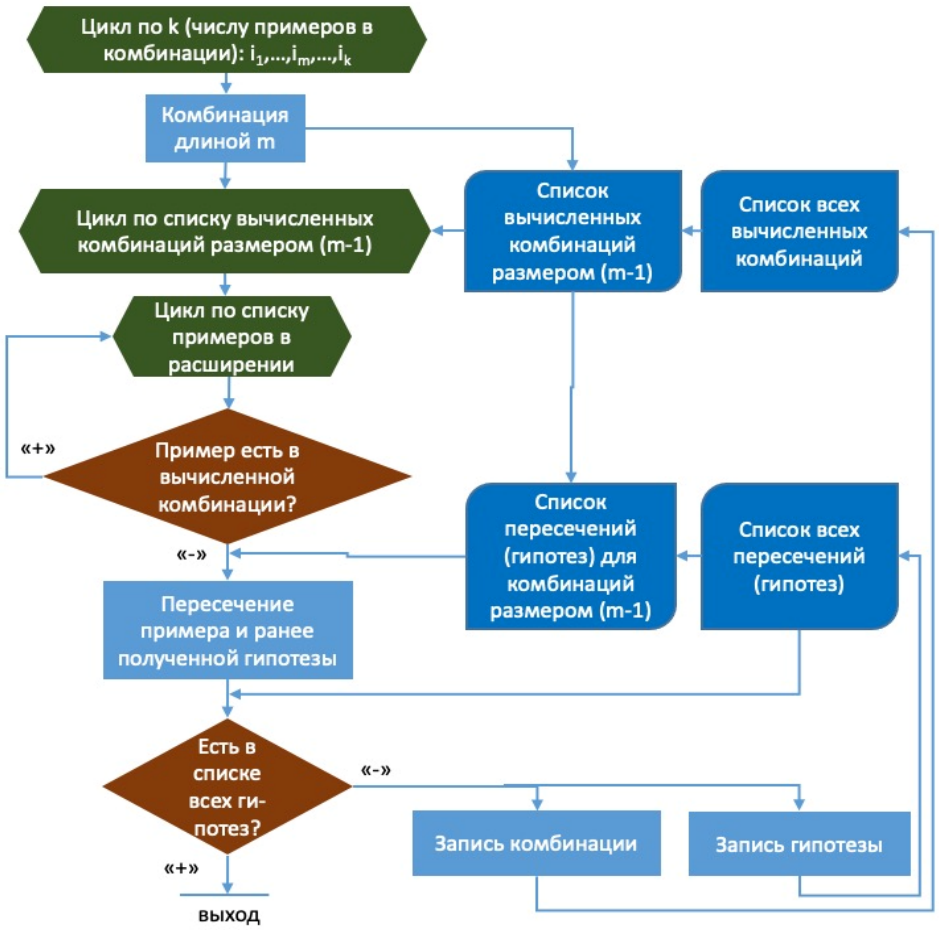

Рис. 3. Схема организации поиска пересечений примеров в рамках одного расширения БФ

В реализованном подходе с *list comprehension* применяются объекты языка *python*, которые относятся к динамическим массивам данных и именуются генераторами (*generator*). Альтернативой этому типу объектов был бы лист, элементами которого являются массивы *tuple*. Однако такой массив необходимо постоянно хранить в памяти, а главное – обращаться к нему с высокой частотой, что неизбежно влечет за собой увеличение времени исполнения программы и ужесточение требований к емкости оперативной памяти. В случае же применения генераторов каждый элемент массива генерируется «на лету», исключительно при обращении к нему, что существенно экономит ресурсы: в частности, в представленной реализации требуется в 1,5 раза меньше времени.

Листинг 1. Псевдокод модуля, осуществляющего порождение гипотез

Объявление пустого словаря (*dictionary*) для хранения гипотез и породивших их комбинаций примеров
Объявление пустого множества (*set*) для хранения всех порожденных гипотез
Цикл: Для каждого расширения в перестановке расширений БФ:
Объявление пустого множества (*set*) для хранения гипотез, порожденных в текущем расширении
Цикл: Для каждого количества пересекаемых примеров:
Если пересекаем два примера:
Цикл: Для каждого индекса примеров в текущем расширении:
Цикл: Для каждого индекса примеров во всей БФ:
Если индексы не равны и представлены в возрастающем порядке:
Нахождение пересечения
Реализация стратегии (пример – стратегия с запретом на контр-примеры):
Цикл: Для каждого примера противоположного знака:
Если порожденная гипотеза содержится в примере противоположного знака:
Выход из цикла
Иначе:
Добавить гипотезу в множество для хранения гипотез, порожденных в текущем расширении
Добавить гипотезу в множество для хранения всех порожденных гипотез
Если размер множества для хранения всех порожденных гипотез изменился (данной гипотезы в нем ранее не было):
Добавить гипотезу в словарь для хранения гипотез и породивших их комбинаций

примеров, вместе с комбинацией породивших ее примеров.

Если пересекаем более чем два примера *(m* примеров):

Цикл: Для каждой комбинации индексов примеров, имеющих размерность *m*-1:

Цикл: Для каждого индекса примеров в текущем расширении:

Если индекс примера присутствует в ранее вычисленной комбинации:

Выход

Иначе:

Нахождение пересечения

Реализация стратегии

Добавить гипотезу в словарь для хранения гипотез и породивших их комбинаций примеров вместе с комбинацией породивших ее примеров

На этапе предсказания неизвестных примеров псевдокод имеет вид, представленный в *Листинге 2*.

Листинг 2. Псевдокод модуля, осуществляющего предсказание неизвестных примеров.

1. Определение τ-примеров (выполняется независимо для каждого знака):

Цикл: Для каждого τ-примера в списке примеров, оставленных на предсказание:

Создание (очистка) рабочих файлов

Цикл: для каждой гипотезы о причинах:

Если гипотеза вкладывается в τ-пример:

В файл, хранящий число вхождений гипотез в τ-пример, в строку с номером текущего τ-примера записывается (прибавляется) единица

Гипотеза записывается в файл гипотез, участвующих в предсказаниях, с ярлыком номера примера и знака

2. Подсчет вхождений гипотез в τ-пример:

Определение знака примера: из файла, хранящего число вхождений гипотез в τ-пример, считывается количество вхождений гипотез каждого знака, сравнивается между собой и делается вывод о предсказанном знаке τ-примера.

Пополняется файл с перечнем (+)-примеров

Пополняется файл с перечнем (-)-примеров

3. Проверка каузальной полноты (выполняется независимо для каждого знака):

Считываются индексы примеров текущего знака

Из БФ считываются наборы признаков для каждого из примеров текущего знака

Цикл: для каждого из примеров текущего знака:

Цикл: для каждой из гипотез о причинах текущего знака:

Если гипотеза вкладывается в пример:

Индекс примера записывается в файл

Цикл по гипотезам прерывается

Иначе:

Цикл по гипотезам продолжается

Иначе:

Пример помечается как необъясненный

Запись в файл протокола исследования списка индексов объясненных и необъясненных примеров

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Представленный способ реализации ДСМ-метода АПИ был выполнен на бытовом компьютере с 6-ядерным процессором и 8 ГБ оперативной памяти, а также на вычислительном кластере *Amazon AWS EC2 Instance r5.8xlarge* с 32 виртуальными ядрами и 256 ГБ оперативной памяти. При запуске на бытовом компьютере параллельное исполнение эксперимента заняло в 4,3 раза меньше времени, чем последовательное (9,2 часа против 39,3 соответственно). В рассматриваемом случае максимальное число процессов, которые могли идти параллельно – шесть, что соответствует числу ядер процессора.

Вычислительный кластер имел 32 ядра процессора, что подразумевает одновременную обработку 32 процессов. При запуске на нем программы время ее выполнения при параллельном варианте оказалось меньше времени выполнения при последовательном в 21,1 раз (1.6 часа против 34.1 соответственно).

ЗАКЛЮЧЕНИЕ

ДСМ-метод автоматизированной поддержки исследований обладает высокой вычислительной сложностью и является достаточно ресурсоемким методом интеллектуального анализа данных по сравнению с большинством алгоритмов машинного обучения, которые для множества задач могут выполняться на бытовых компьютерах. Без оптимизации ДСМ-метод АПИ требует значительное количество вычислительных ресурсов, что может быть критично в условиях ограниченного времени исследователя либо недостатка техники или мощностей. Однако современные вычислительные станции, имеющие многоядерную архитектуру процессоров, способны нивелировать данный недостаток при условии параллельной реализации ДСМ-метода АПИ, пример которой был продемонстрирован в нашей работе.

Из результатов эксперимента видно, что число ядер процессора и время эксперимента связаны нелинейной зависимостью. Иными словами, шесть параллельных процессов не дает шестикратную экономию времени на бытовом компьютере, и аналогично 23 процесса не сокращают время вычисления в 32 раза. В рассматриваемом случае этот факт связан с тем, что не все стратегии имеют одинаковую сложность выполнения. Так, стратегия различия занимает наибольшее время расчета, которое примерно в два раза превышает время, необходимое для остальных стратегий. Одно из направлений будущего совершенствования системы – оптимизация алгоритма реализации стратегии различия.

Другое немаловажное направление оптимизации вычислительного времени ДСМ-метода АПИ – это экономия ресурсов оперативной памяти, которая также может быть выполнена путем использования таких типов данных, которые требуют меньше места для хранения.

Серьезным барьером на пути внедрения ДСМ-метода АПИ в различные области науки и промышленности является отсутствие стандартных библиотек в популярных языках программирования. Действительно, многие методы машинного обучения могут

быть легко применены исследователями, не обладающими глубокими знаниями программирования, при помощи вызова стандартных функций, зачастую с предустановленными начальными параметрами. Для ДСМ-метода АПИ существуют программные инструменты, однако они созданы для решения определенного класса задач, и не могут быть легко кастомизированы под нужды задач иного типа [6].

Одно из решений проблемы – реализация ДСМ-метода АПИ собственными силами исследователя, при условии, что он владеет азами программирования на любом из языков, т. е. принадлежит к той же аудитории, которая способна применять алгоритмы машинного обучения при помощи библиотек, например *scikit-learn*. В таком случае следует ожидать повышение распространенности ДСМ-метода АПИ в различных отраслях науки.

Надеемся, что работа поможет исследователям из различных сфер в реализации решателя ДСМ-метода АПИ как на *python*, так и на других языках программирования.

СПИСОК ЛИТЕРАТУРЫ

1. Чебанов Д.К., Михайлова И.Н. Интеллектуальный анализ данных пациентов с меланомой для поиска маркеров заболевания и значимых генов // Научно техническая информация. Сер. 2. – 2019. – № 10 – С. 35-40.
2. Финн В.К. Об эвристиках ДСМ-исследований (дополнения к статьям) // Научно-техническая информация. Сер. 2. – 2019. – № 10. – С. 1-34.
3. Финн В.К. Об определении эмпирических закономерностей посредством ДСМ - метода автоматического порождения гипотез // Искусственный интеллект и принятие решений. – 2010. – № 4. – С. 41-48.
4. Финн В.К. Дистрибутивные решетки индуктивных ДСМ-процедур // Научно-техническая информация. Сер. 2. – 2014. – № 11. – С. 1-36
5. ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания / сост. О.М. Аншаков, Е.Ф. Фабрикантова; под. общ. ред. О.М. Аншакова. – М.: ЛИБРОКОМ, 2009. – 433 с.
6. Шестерникова О.П., Финн В.К., Винокурова Л.В., Лесько К.А., Варварина Г.Г., Тюляева Е.Ю. Интеллектуальная система для диагностики заболеваний поджелудочной железы // Научно техническая информация. Сер. 2. – 2019. – № 10. – С. 41-48.
7. Kuznetsov S.O., Obiedkov S.A. Comparing performance of algorithms for generating concept lattices // Journal of Experimental and Theoretical Artificial Intelligence. – 2002. – Vol. 14. – P. 189-216.

Материал поступил в редакцию 25.05.2020

Сведения об авторах

ЧЕБАНОВ Дмитрий Константинович – генеральный директор ООО «ОнкоЮнайт Клиникс», Москва
e-mail: chebanov.dk@gmail.com

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 004.93:070

Ал-др А. Хорошилов, Р.Р. Мусабаев, Я.Д. Козловская, Ю.В. Никитин, А.А. Хорошилов

Автоматическое выявление и классификация информационных событий в текстах СМИ*

Описывается решение проблемы автоматического выявления и классификации информационных событий в текстах СМИ на основе модели фразеологического концептуального анализа текстов. Предлагаемое решение базируется на использовании ранее разработанных методов формализации смысловой структуры предложений, а также на методах и алгоритмах установления фрагментов текстов СМИ, описывающих информационные события. В разработанном алгоритме реализованы правила падежной грамматики Ч. Филлмора, базирующиеся на процедурах семантико-синтаксического и концептуального анализа текстов.

Ключевые слова: *выявление информационных событий, классификация информационных событий, семантико-синтаксический анализ текстов, концептуальный анализ текстов, наименования понятий, статистическая мера значимых наименований понятий, семантический корреляционный коэффициент, меры смысловой близости содержания текстов и рубрик классификатора*

DOI: 10.36535/0548-0027-2020-07-4

ВВЕДЕНИЕ

Современное общество в процессе своего развития порождает огромные объемы текстовой информации, в том числе так называемой новостной, генерацию которой осуществляют средства массовой информации (СМИ). Результатом их деятельности является информирование общества о событиях[1], происходящих в реальном мире.

Некоторое время назад огромную популярность приобрели электронные средства массовой информации (ЭСМИ). Они являются одним из наиболее эффективных способов воздействия на психоэмоциональный настрой современного общества и поэтому часто используются в качестве инструмента в современной информационной войне. Некоторые ЭСМИ наряду с объективным освещением событий нередко подают информацию в извращенном виде. Это делается для того, чтобы сформировать у определенной части социума заранее заданный психоэмоциональный настрой. Для борьбы с такими манипуляциями при освещении информационных событий необходимо выработать автоматизированные средства противодействия, базирующиеся на методах и технологиях смыслового анализа текстов.

Поэтому автоматическое выявление в текстовом потоке новостной информации описаний информационных событий (инфоповодов)[2], формализация их смыслового содержания и классификация являются крайне важной научной и методологической задачей. Современные средства автоматической обработки и анализа текстовой информации пока не достигли такого уровня зрелости, который позволил бы в полной степени ее реализовать. Одним из возможных путей решений этой проблемы являлась бы разработка средств автоматической формализации и анализа смыслового содержания различных источников информации (ЭСМИ, социальные сети, теле и радиове-

* Статья подготовлена в рамках проекта ПЦФ BR05236839 «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана».

[1] Под термином *событие* (*информационное событие)* будем понимать описания в сообщениях СМИ социально значимых явлений, происшествий, фактов общественной деятельности мирового или регионального масштаба, а также факты и события деятельности социальных конгломераций или факты личной жизни известных общественных деятелей и др.

[2] Под термином *инфоповод* понимается главная тема сообщения, заставляющая целевую аудиторию его обсуждать. Как правило, информационный повод отражает важные факты содержания события.

щание и др.). В процессе анализа текстов СМИ необходимо производить автоматическое извлечение фактологической информации, ее структурирование и классификацию по заданному перечню актуальных проблем, а также создавать на основе структурированной информации фактологическую базу знаний, которая позволила бы оперативно решать ряд информационных и аналитических задач. Но создание такой базы знаний – это трудоемкий процесс, требующий участия высококвалифицированных специалистов. Основным технологическим процессом разработки и ведения баз знаний является извлечение фактографической информации из новостного потока неструктурированной текстовой информации и формализация ее смыслового содержания. Автоматизация этого процесса является достаточно сложной задачей, требующей использования современных программных средств интеллектуальной обработки текстовой информации.

СОВРЕМЕННОЕ СОСТОЯНИЕ ТЕХНОЛОГИЙ ИЗВЛЕЧЕНИЯ ФАКТОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

В настоящее время разработан ряд технологий извлечения фактологической информации (ИФИ) из текстов на естественном языке, позволяющих автоматически выявлять целевые факты, объекты и отношения между ними в виде, пригодном для дальнейшей автоматической обработки: статистической обработки, визуализации, структурирования и поиска закономерностей в данных и др. [1–4].

Одним из наиболее распространенных методов ИФИ, является метод, базирующийся на использовании концептуальных графов и решетки понятий [1]. Идея этого метода состоит в том, что на предложениях обрабатываемых текстов строится множество концептуальных графов и решается задача их агрегирования для исключения избыточной размерности концептуальных моделей. Для этого создается формальный контекст, задаваемый матрицей отношений на множествах объектов и их атрибутов. В завершении выделяются формальные понятия и строится графовая модель вида решетки понятий, позволяющая выявлять связи между ними как «общее-частное». Описанный метод позволяет выявлять в текстах именованные сущности и отношения между ними.

Промышленные решения выделения фактографической информации разработаны в компании *RCO*. Они базируются на преобразовании текста в его представление в виде семантической сети и трансформации описаний фактов в виде их семантико-синтаксических шаблонов. В процессе построения семантической сети документа распознаются особые текстовые конструкции (паспортные и регистрационные данные фигурантов, адреса, телефоны, даты и др. конструкции) и определяются лексико-грамматические характеристики элементов текста. При поиске шаблонов фактов в его узлах и связях при помощи логических выражений указываются условия, которым должны удовлетворять эти узлы и связи семантической сети документа. Ключевой особенностью текста досье [3] является высокая плотность связей между словами, которые выражаются средствами анафорических связей. Досье часто содержит семантически сложные предложения, требующие для их анализа привлечения полного арсенала средств компьютерной лингвистики.

Технологии, используемые в системе ИСИДА-Т [2] требуют настройки на предметную область или на решение конкретной задачи извлечения информации, в которой задается искомый факт. В основу этих технологий заложены:

- средства определения кодировки документа, извлечения текстового слоя, его предварительной фильтрации;
- разделение текста на отдельные слова, морфологический и синтаксический анализ, определение границ предложений;
- осуществление поиска в документе целевой лексики и синтаксических конструкций, а также первичное структурирование информации, отождествление элементов знаний, вывод производных знаний и приведение извлеченной информации к определенному формату.

Компанией «Яндекс» был разработан инструмент извлечения фактологической информации – Томита-парсер [2]. Этот инструмент создан на основе алгоритма *GLR*-парсинга. Он позволяет выделять из текста по заранее заданным правилам цепочки слов, содержащих фактологическую информацию. Грамматика для Томита-парсера представляется в виде множества правил на языке контекстно-свободных грамматик, позволяющих описывать структуру выделяемых цепочек и словарей ключевых слов, представляющих набор статей, состоящих из множества слов и словосочетаний, объединённых общим свойством. Результатом работы Томита-парсера являются списки словосочетаний из текста, полученные в соответствии с используемой грамматикой. При этом слова в словосочетании остаются как в согласованном виде, так и по возможности приводятся к нормальной форме.

Некоторые технологии извлечения фактологической информации (ИФИ) используют существующие онтологические ресурсы, например, онтологии *WordNet* [4, 5] или такой ресурс как *Wikipedia*. Методы оценки систем ИФИ также нуждаются в доработке, т.к. большинство из них основываются на их сравнении с эталонной разметкой корпуса текстов. Наименее разработанными методами являются методы ИФИ, обеспечивающие возможность обработки неструктурированной текстовой информации по произвольной тематике.

Несмотря на разнообразие существующих методов ИФИ в них до сих пор не решены ключевые проблемы смыслового анализа текстов. Исходя из вышеизложенного, можно сделать вывод, что для адекватного выявления фактографической информации из текстов документов необходимы методы ИФИ, базирующиеся на современных теоретических представлениях о смысловой структуре текстов и методах интеллектуальной обработки текстовой информации. В качестве одного из таких представлений можно использовать модель фразеологического концептуального анализа текстов и разработанные на ее основе методы семантико-синтаксического и концептуального анализа текстов.

КОНЦЕПЦИЯ СМЫСЛОВОЙ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ

В основу используемой концепции смысловой обработки текстовой информации положена модель фразеологического концептуального анализа текстов. Она базируется на предположении, что содержание текстов выражается через систему его базовых единиц смысла – фразеологических и терминологических понятий (сущностей) и их смысловых отношений. С помощью таких понятий формируются смысловые единицы более высоких уровней: предложения и сверхфразовые единства.

Предложения также являются значимыми единицами смысла. Их основное свойство – предикативность, т.е. наличие у объектов определенных признаков и их отношений, выражающихся в текстах через предикатно-актантную структуру предложений. Ее компонентами являются понятия-предикаты (признаки и отношения) и понятия-актанты, выступающие в роли описываемых объектов. Выявление в текстах формальной предикатно-актантной структуры выполняется на этапах семантико-синтаксического и концептуально анализа текстов [6]. При формализации смысловой структуры фактологической информации необходимо не только выделить компоненты предикатно-актантной структуры, но и связать их ролевыми функциями в предложении. Здесь необходимо обратиться к падежной (ролевой) грамматике Ч. Филлмора [7].

В соответствии с этой грамматикой Ч. Филлмор представил план содержания предложения как препозицию – вневременной набор семантико-синтаксических функций («ролей») для именных компонентов, чей состав и взаимоотношения задаются лексическим значением глагольного знака. Филлмор определяет семантические роли как «концепты, в терминах которых человек судит о происходящих вокруг него событиях». При этом каждому глаголу должен соответствовать определенный набор падежей (падежный фрейм). Каждому падежу соответствует определенный участник (партиципант) события. Совокупность черт, характерных для одинаково кодируемых партиципант, называется семантической ролью.

Филлмор выделил следующие основные семантические роли:

1) A – агентив (субъект действия); (от латинского слова agens 'действующий': партиципант, осуществляющий контроль над ситуацией; тот, по чьей инициативе она разворачивается);
2) *D* – датив (лицо, затронутое объектом действия);
3) *T* – инструменталь (инструмент или действие);
4) *F* – фактив (результат действия);
5) *L* – локатив (пространство действия);
6) *O* – объектив (объект действия);
7) *P* – пациенс (партиципант, на которого направлено воздействие и чьё физическое состояние, в том числе положение в пространстве, изменяется в результате осуществления этой ситуации);
8) *E* – экспериенцер (партиципант, на чье внутреннее состояние ситуация оказывает воздействие);
9) *S* – стимул (партиципант, который является источником воздействия, оказываемого на внутреннее состояние другого участника ситуации);
10) *R* – реципиент (партиципант, приобретающий что-то в ходе реализации ситуации).

Необходимо отметить, что в основе постулируемого набора семантических ролей лежит не классификация самих парципантов, а классификация возможных их наборов, т.е. семантических фреймов. Считается, что семантические роли позволяют при анализе предложения учитывать его глубинную структуру на основе предварительно заданной «модели мира» в терминах семантических ролей.

Таким образом, исходя из концепции Филлмора, можно констатировать, что семантическая структура предложения определяется как препозиция – «… вневременной набор отношений между глаголом (предикатом) и именами (концептами)». Отношение между предикатом и концептом называется семантическим (или глубинным) падежом, значение которого вскрывается на основе их трансформаций (преобразований). Семантические падежи элементарны и дальнейшему анализу не подлежат. Каждый падеж входит в структуру высказывания только один раз. Падежные отношения отображают базисные мыслительные универсалии, существующие между участниками реальных ситуаций.

В рамках этой грамматики предлагаются правила перехода от глубинных структур к их поверхностным реализациям:

1) правила токенизации (выбор класса высказывания);
2) правила субъективизации (установление синтаксического подлежащего);
3) правила объективизации (установление синтаксических дополнений);
4) правила установления падежной формы (определение формы прямого или косвенного дополнения).

В процессе смысловой обработки текстовой информации, базирующейся на процедурах семантико-синтаксического и концептуального анализа текстов, возможно реализовать вышеприведенные правила падежной грамматики. Эти процедуры должны опираться на адекватные семантико-синтаксические модели текстов, в которых понятия представляются преимущественно фразеологическими словосочетаниями. В ходе такого анализа необходимо определить синтаксическую структуру предложений текста, выявить систему его понятий и установить смысловые отношения между элементами этой системы.

Проблема унификации различных текстовых форм наименований понятий, выражающих одинаковый смысл, снимается путем их трансформации в унифицированное формализованное представление [6, 8, 9].

МЕТОДЫ ВЫЯВЛЕНИЯ НАИМЕНОВАНИЙ ПОНЯТИЙ В ТЕКСТАХ

Извлечение информации связано, прежде всего, с поиском значимых наименований понятий и их смысловых отношений. Это один из ключевых этапов предварительной обработки текста, необходимый для

реализации более сложных моделей извлечения фактов. Сложность выявления наименований понятий, представленных словосочетаниями, заключается в правильном установлении их границ в текстах и выявлении тех наименований понятий, которые в тексте несут основную смысловую нагрузку. Эти проблемы решаются методами статистического, синтаксического и концептуального анализа текстов.

Статистические методы позволяют путем назначения весовых коэффициентов установить состав значимых слов и словосочетаний на основе анализа их частот в конкретном тексте и в тематическом корпусе текстов [10]. Синтаксические методы дают возможность выявить синтаксическую роль значимых слов и словосочетаний путем установления их синтаксической роли в предложении – принадлежность к словосочетаниям, являющимся в предложении группой подлежащего (субъекта), группой сказуемого (предиката) или группой дополнения, обстоятельств места и времени (объектов) [6]. При помощи семантических методов в тексте выявляются значимые слова и словосочетания путем их соотнесения с элементами эталонных словарей.

Для реализации этой задачи в тексте необходимо установить систему наименований понятий и назначить им весовые коэффициенты смысловой значимости [10]. В качестве меры смысловой значимости слов и словосочетаний часто используется так называемая статистическая мера *TF-IDF* (*TF* (*term frequency*) – частота слова.

TF – отношение числа вхождений наименования понятия общему числу наименований понятий документа. Таким образом, оценивается важность наименования понятия t_i в пределах отдельного документа:

$$tf\left(t,d\right)=\frac{n_t}{\sum_k n_k}\ , \qquad (1)$$

где n_t – число вхождений наименования понятия *t* в документ;

$\sum_k n_k$ – общее число наименований понятий в данном документе.

IDF – инверсия частоты, с которой некоторое наименование понятия встречается в документах коллекции. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение IDF:

$$idf\left(t,D\right)=\log\frac{\left|D\right|}{\left|\left\{d_i\in D\middle|t\in d_i\right\}\right|}\ , \qquad (2)$$

где: $\left|D\right|$ – число документов в коллекции; — число документов из коллекции *D*, в которых встречается *t* (когда $n_t\neq 0$).

Мера *TF-IDF* является произведением двух сомножителей:

$$tf-idf\left(t,d,D\right)=tf\left(t,d\right)\times idf\left(t,D\right). \qquad (3)$$

Большой вес в *TF-IDF* получат наименования понятий с высокой частотой встречаемости в конкретном документе и с относительно низкой частотой в пределах всего корпуса текстов.

Но такая статистическая мера в явном виде не отражает смысловую составляющую наименований понятий. С этой целью была разработана система коррелирующих семантических весовых коэффициентов наименований понятий восполняющая этот пробел. Обобщенная мера смысловой значимости наименований понятий M_7 с учетом этих коэффициентов вычисляется по формуле:

$$M_7=\left(TF\cdot IDF\right)\cdot K_n\cdot K_z\cdot K_t\cdot K_w\cdot K_s\cdot K_f \qquad (4)$$

где: K_n – коэффициент, учитывающий распознающую способность слов при их нормализации;

K_z – коэффициент, учитывающий вхождение в заголовки слов или словосочетаний;

K_t – коэффициент, учитывающий вхождение в эталонный концептуальный словарь или другой семантический инструмент;

K_w – коэффициент, учитывающий количество слов в словосочетании;

K_s – коэффициент, учитывающий синтаксическую роль слова или словосочетания в предложении;

K_f – коэффициент, учитывающий принадлежность понятия к фамильно-именной группе, бренду и др.

ИЗВЛЕЧЕНИЯ И ФОРМАЛИЗАЦИЯ ИНФОРМАЦИОННЫХ СОБЫТИЙ НА ОСНОВЕ СИНТАКСИЧЕСКОЙ МОДЕЛИ ТЕКСТА

Как было описано в работе [6], реализованная нами синтаксическая модель текста на основе системы обобщенных синтагм позволяет соотнести его смысловую и синтаксическую структуру. При использовании этой модели можно формальным образом выявить понятийный состав текста и установить смысловые отношения между понятиями (сущностями). Установление их синтаксической роли в предложении позволяет определить в тексте основные синтаксические конструкции представления смысловой структуры предложения. В тех случаях, когда в роли «субъектов», «предикатов» или «объектов» предложений выступают значимые именованные сущности, такие слова и словосочетания фиксируются как элементы формального описания событий. Далее эти элементы классифицируются в соответствии с их ролевыми функциями, приводятся к их унифицированным формам представления и структурируются в соответствии с системными требованиями.

Важной задачей такого процесса является установление границ описания информационного события, представленного в тексте контактно расположенной последовательностью предложений. При этом необходимо исходить из того факта, что предложения выступают в тексте не изолированно друг от друга, а находятся в тесной смысловой связи. В ее основе лежат мыслительные образы тех конкретных или

абстрактных объектов (ситуаций, явлений), которые человек имеет в виду, когда он порождает текст. Образы этих объектов имеют определенную структуру. Кроме того, они дополнительно структурируются человеком при их описании на естественном языке.

В процессе описания событий текст развертывается последовательно, т.е. имеет линейную структуру, тогда как мыслительные образы этого события – "многомерны". При их описании может быть принят различный порядок линейной развертки, но цель описания должна быть в основном одна и та же – воссоздание в сознании читателей мыслительных образов, подобных мыслительным образам автора текста. Такое воссоздание осуществляется постепенно, путем восприятия предложения за предложением и "монтажа" возникающих при этом частичных образов в целостный мыслительный образ, соответствующий содержанию текста. При этом в каждом предложении элемент его актуального членения – "тема"– выполняет роль "стыковочного узла", служащего для подключения нового частичного мыслительного образа, обозначаемого этим предложением, к ранее построенному мыслительному образу.

Описанная модель восприятия текста позволяет объяснить тот факт, что связи между предложениями выражаются в большинстве случаев с помощью лексических повторов: в "стыковочных узлах" предложений повторяются наименования понятий предшествующего текста либо буквально, либо в виде синонимических и эллиптических конструкций, либо в виде родовых наименований понятий и местоимений. Соответственно при установлении границ описания информационного события необходимо выявить в тексте цепочки смысловых связей наименований понятий и установить способы выражения смысловой связи между предложениями. Такими способами могут быть [9]:

- повтор слов и словосочетаний. При этом это может быть либо буквальное совпадение единиц текста, либо совпадение с точностью до словоизменения, либо совпадение с точностью до словообразования;
- синонимия слов и словосочетаний, включая синонимию, связанную с аббревиацией;
- родо-видовые отношения между показателями связи. При этом предшествующий по тексту показатель связи может обозначать видовое понятие, а последующий – родовое;
- эллипсис – окказиональное опущение отдельных слов или групп слов в словосочетаниях, выступающих в роли показателей связи;
- местоименная анафора – замена слова или словосочетания предшествующего предложения на замещающее его местоимение;
- другие средства выражения межфразовой связи (в частности, вводные слова, союзы и наречия).

Таким образом, описания конкретных событий в тексте могут отображаться его контактно расположенной последовательностью предложений, соединенных, межфразовыми связями. Рассмотрим выражения межфразовой связи между предложением на конкретном примере. Возьмем следующий текст СМИ, представленный в табл. 1.

Для последующего анализа необходимо расчленить текст на отдельные предложения и перенумеровать их, а также присвоить каждому предложению идентификационный индекс, состоящий из порядкового номера события, номера предложения в нем, источника информации и даты публикации (табл. 2).

В этом фрагменте текста межфразовая связь между предложениями № 1 и № 2 обусловлена местоименной анафорой ***«об этом»,*** связь между предложениями № 1 и № 3 – синонимичными конструкциями «***НАТО – Североатлантический альянс***», связь между предложениями № 3 и № 4 – смысловой связью между наименованиями понятий «***странах ЕС вдоль западной границы России: в Латвии, Литве, Эстонии, Польше, Болгарии и Румынии - государств Восточной Европы***», связь между предложением № 5 и предшествующими предложениями № 1 – №4 обусловлена вводным словом предложения № 5 «кроме того». Таким образом, этот фрагмент текста состоит из предложений, соединенных выше рассмотренными межфразовыми связями, что является необходимым условием выделения описания конкретного информационного события в тексте. Дополнительным условием, представленным в табл. 3, выступает смысловая связь (родовидовые и синонимичные отношения) между частью наименований понятий этого фрагмента, находящихся в разных предложениях (в скобках указаны номера предложений).

Используя формализацию смыслового представления понятий и рассмотренные выше межфразовые связи, а также смысловые отношения между элементами предложений, можно построить различные конструкции синтаксически и семантически связанных наименований понятий предложений рассматриваемого фрагмента в соответствии с классификацией Ч. Филлмора. В табл. 4 на первой позиции указывается мнемоника состава понятий падежной грамматики предложений, сопровождаемая представлением семантических ролей наименований понятий. При этом, поскольку предложение № 2 связано с предыдущим предложением, но не содержит в своем составе значимых наименований понятий, его мы исключаем из дальнейшего анализа.

В рамках разработанной нами синтаксической модели возможно осуществить автоматическое преобразование исходного текста в его формализованную синтаксическую структуру, аналогичную приведенной выше. В процессе такого преобразования для получения унифицированного представления элементов предикатно-актантной структуры (ПАС) предложения необходимо дополнительно также выполнить нормализацию форм слов наименований понятий и их унификацию по словарю унифицированных формализованных представлений наименований понятий. В результате этой операции получим список наименований понятий, приведенных к их унифицированным формам представления (табл. 5).

Теперь, если нормализовать и унифицировать наименования понятий из табл. 4, то она приобретет иной вид (табл. 6). В этом представлении будем применять ранее используемую мнемонику, принятую в работах [6, 8, 9].

Таблица 1

Фрагмент текста СМИ

Информационное событие (Information event) №2671от 07.12.2019 (IE-2671-07.12.19))

"……НАТО перебрасывает войска в соседние с Россией страны. Об этом мы писали вчера. По программе «Расширенные передовые силы» Североатлантический альянс создает и укрепляет военные группировки в странах ЕС вдоль западной границы России: в Латвии, Литве, Эстонии, Польше, Болгарии и Румынии. В каждом из государств Восточной Европы разместят около полутора тысяч военных и сотни танков, артиллерийских установок и прочей военной техники. Кроме того, в Черном море НАТО создало минигруппу флота, в которую входят не только причерноморские страны, но еще и Польша, Испания и Германия……»

Таблица 2

Перечень контактно-расположенных предложений, связанных межфразовыми связями

1. НАТО перебрасывает войска в соседние с Россией страны ***-IE-2671_173_1-07.12.19****.*

2. Об этом мы писали вчера ***-IE-2671_173_1-07.12.19****.*

3. По программе «Расширенные передовые силы» Североатлантический альянс создает и укрепляет военные группировки в странах ЕС вдоль западной границы России: в Латвии, Литве, Эстонии, Польше, Болгарии и Румынии ***-IE-2671_173_1-07.12.19****.*

4. В каждом из государств Восточной Европы разместят около полутора тысяч военных и сотни танков, артиллерийских установок и прочей военной техники ***-IE-2671_173_1-07.12.19****.*

5. Кроме того, в Черном море НАТО создало минигруппу флота, в которую входят не только причерноморские страны, но еще и Польша, Испания и Германия ***-IE-2671_173_1-07.12.19****.*

Таблица 3

Списки синонимичных и родо-видовых понятий, с указанием их местоположений в предложениях текста

- *НАТО* ***(№1, №5)****-Североатлантический альянс* ***(№2)****-(в словаре УФПНП*[3] *=страны НАТО: Латвия, Литва, Эстония, Польша, Болгария, Румыния, Испания, Германия* ***(№3,№5)****;*
- *войска****(№1)****-военные группировки* ***(№3)****-минигруппу флота* ***(№5)****;*
- *страна НАТО-соседние с Россией страны* ***(№1)****- странах ЕС вдоль западной границы России* ***(№3)****, Латвия, Литва, Эстония, Польша, Болгария, Румыния* ***(№3)****-Польша, Испания и Германия* ***(№5)****.*

Таблица 4

Формализованное представление предложений в мнемонике грамматики Ч. Филлмора

1. PAOL=*(что делает?) P=перебрасывает; (кто?) A=НАТО; (что?) O=войска; (где?) L=в соседние с Россией страны)* ***-IE-2671_1-173_07.12.19****..*

2. PAOLD=*(что делает?) P=создает; (кто?) A=Североатлантический альянс; (что?) O=военные группировки; (где?) L=в странах ЕС вдоль западной границы России; (где?) L=в Латвии, Литве, Эстонии, Польше, Болгарии и Румынии; (по чему?) D=по программе «Расширенные передовые силы»* ***-IE-2671_3-173_07.12.19****.*

3. PAOLD=*(что делает?) P=укрепляет; (кто?) A=Североатлантический альянс; (что?) O=военные группировки; (где?) L=в странах ЕС вдоль западной границы России:-в Латвии, Литве, Эстонии, Польше, Болгарии и Румынии; (по чему?) D= по программе «Расширенные передовые силы»* ***-IE-2671_173_3-07.12.19****.*

4. POL=*(что делает?) P=разместят; (что?) O=около полутора тысяч военных и сотни танков, артиллерийских установок и прочей военной техники; (где?) L=в каждом из государств Восточной Европы****-IE-2671_173_4-07.12.19****.*

5. PAOL=*(что делает?) P=создало; (кто?) A=НАТО; (что?) O=минигруппу флота; (где?) L=в Черном море****-IE-2671_173_5-07.12.19****.*

6. PAO=*(что делает?) P=входят; (кто?) A=НАТО; (что?) O=не только причерноморские страны, но еще и Польша, Испания и Германия; (где?)* ***-IE-2671_173_5-07.12.19****.*

[3] Словарь унифицированных формализованных представлений наименований понятий (УФПНП) [2]

Таблица 5

Список унифицированных форм представления наименований понятий

- *государство -> страна,*
- *североатлантический альянс->НАТО,*
- *военные группировки-> войско,*
- *минигруппу флота-> войско,*
- *странах ЕС вдоль западной границы России -> страна,*
- *государств Восточной Европы -> страна,*
- *причерноморские страны ->страна,*
- *Латвия-> страна,*
- *Литва-> страна,*
- *Эстония-> страна,*
- *Польша-> страна,*
- *Болгария-> страна,*
- *Румыния-> страна,*
- *Испания-> страна,*
- *Германия -> страна.*

Таблица 6

Формализованное представление предикатно-актантной структуры предложений

PSOL# *P=перебрасывать; S=НАТО; O= войско; L=страна* ***-IE-2671_173_1-07.12.19***
PSOLD# *P=создать; S=НАТО; O= войска; L=страна; D=программа****-IE-2671_173_1-07.12.19****;*
PSOL# *P=укрепить; S=НАТО; O= войска; L=страна* ***-IE-2671_173_1-07.12.19***
PSOL# *P=разместить; O= войск; L=страна* ***-IE-2671_1-173_07.12.19***
PSOL# *P=создать; O= войско; L=море; L=страна* ***-IE-2671_173_1-07.12.19***

На основе вышеприведенного анализа концепции падежной грамматики Ч.Филлмора и правил ее реализации, а также базируясь на синтаксической модели представления смысловой структуры текстов, мы разработали алгоритм автоматического выявления и формализации информационных событий в текстах СМИ.

Алгоритм. Автоматическое выявление и формализация информационных событий в текстах СМИ.

Шаг 1. Обработать анализируемый документ процедурой деления на предложения и выполнить обработку каждого из них процедурой морфологического анализа.

Шаг 2. Выполнить упрощенный семантико-синтаксический анализ каждого предложения текста, установить в анализируемом предложении именные и глагольные словосочетания и определить их синтаксическую роль в предложении.

Шаг 3. Выполнить концептуальный анализ текста. Составить частотный словарь слов и словосочетаний.

Шаг 4. На основе информации о статистических, синтаксических и семантических параметрах словосочетаний определить для текста список семантически значимых слов и словосочетаний.

Шаг 5. Каждому словосочетанию присвоить по словарю унифицированных формализованных представлений наименований понятий его уникальный идентификатор и соотнести исходную форму слова или словосочетания с их унифицированными формами представления, установить их местоположение в тексте (номера предложений в тексте).

Шаг 6. По словарю указателей связей предложений установить их местоположение в тексте (номера предложений в тексте).

Шаг 7. Установить межфразовые связи между предложениями, содержащими значимые наименования понятий и окружающими их предложениями. Используя разметку текста по составу обобщенных наименований понятий и унифицированных указателей смысловых связей между предложениями установить границы описания событий в тексте.

Шаг 8. Каждому из текстовых описаний событий присвоить идентификационный номер, состоящий из порядкового номера события, кода источника документа и даты публикации.

Шаг 9. На основе результатов выполненного упрощенного синтаксического анализа предложения на шаге 2 определить границы словосочетаний, главные и второстепенные члены предложения, построить дерево зависимостей предложения, построить предикатно-актантную структуру и сформировать «скелет» предложения.

Шаг 10. Для каждого словосочетания определить позицию в предложении и длину, установить главное слово и построить формализованное представление, установить «внешнее» управление (его «хозяин» в предложении) и построить обобщенную синтагму.

Шаг 11. Соотнести полученные словосочетания методом концептуального анализа со словосочетаниями, полученными путем синтаксического анализа.

Шаг 12. Произвести автоматическую нормализацию каждого словосочетания предложения.

Шаг 13. Расчленить описание каждого события на составные элементы – формализованное представление элементов предикатно-актантной структуры, «скелет» предложения с указанием номеров словосочетаний в динамически пополняемом эталонным концептуальном словаре, а также представить само предложение в виде последовательности синтагм и нормальных форм слов предложения.

Шаг 14. Выполнить генерацию формализованных представлений предложений информационного события в обобщенное формализованное представление его смысловой структуры.

Шаг 15. Произвести преобразование обобщенного формализованного представления в его машинную форму. Входом в словарную статью служит перечень унифицированных элементов предикатно-актантной структуры события.

В табл. 7 приведены промежуточные результаты работы алгоритма выявления и формализации событий в текстах сообщений СМИ. В них можно увидеть результаты обработки каждого предложения процедурами семантико-синтаксического и концептуального анализа текстов. Необходимо отметить, что в отличие от уже существующих моделей в рамках предлагаемой модели смысловое представление текста – это иерархия синтаксических конструкций единиц смысла. При этом каждая из них состоит из формы представления единицы смысла и его содержания.

Под формой единицы смысла понимается представление его в виде обобщенной синтагмы, являющейся обобщенным представлением форм слов в их контекстном окружении в виде последовательности их синтагм. Под содержанием единицы смысла понимается формализованное представление смысла в виде нормализованных или унифицированных форм представлений единиц смысла: наименований понятий, предложений-высказываний и сверхфразовых единств.

Таблица 7

Результаты семантико-синтаксического и концептуального анализа предложений описания информационного события, приведенного в табл. 1.

Исходное предложение №1 инфоповода-IE-2671-173_07.12.19	
НАТО перебрасывает войска в соседние с Россией страны	
1.1.1 Формализованное описание элементов ПАС предложения -IE-2671_173_1-07.12.19 Type=PSOL={V,N,N,N}={перебрасывать;нато;войско,страна}	
1.1.1 Формализованное представление элементов предикатно-актантной структуры предложений	
Predicate (P)	*ЮО=перебрасывать#ЮО=перебрасывать*
Subject (S)	*щА=нато#щА= нато*
Object (O)	*ЖВ=войско#ЖВ= войско*
Location (L)	*иq =страна #7АЦТ5АхКиq = в_соседний_с_россия_страна*
1.1.2 Формализованное представления предикатно-актантной структуры предложений (PSORus)	
PSOL=ЮОщАЖВиq*=перебрасывать;НАТО;войско;страна # щАЮОЖВ7Аиq*	
1.1.3 Формализованное представления «скелета» предложений (SklRus)	
щАЮОЖВ7Аиq=*нато_0658341;перебрасывать_0865383;войско_0378327;в;страна_09334562*	
1.1.4 Формализованное представления предложений (SenRus)	
щАЮОЖВ7АЦТ5АхКиq=*нато_перебрасывать_войско_в_соседний_с_россия_страна*	
Исходное предложение №3 инфоповода-IE-2671-173_07.12.19	
По программе «Расширенные передовые силы» Североатлантический альянс создает и укрепляет военные группировки в странах ЕС вдоль западной границы России: в Латвии, Литве, Эстонии, Польше, Болгарии и Румынии.	
1.3.1 Формализованное описание элементов предикатно-актантной структуры предложения -IE-2671_173_3-07.12.19 Type=PSOL={V,N,N,FN}={создать,укрепить;нато;войско;страна}	
Predicate (P)	*ЮО =создать# ЮО= создать* *ЮО =укрепить# ЮО = укрепить*
Subject (S)	*АА = альянс # ЦUАА = североатлантический альянс*
Object (O)	*wS= группировка# ФrwS = военный группировка*
Location (L)	*иЕ=страна#7АиЕшАяАФhДqxS7АхSuGxSvGxSэАхS=в_страна_ЕС_вдоль_западный_граница_Россия в Латвия, Литва, Эстония, Польша, Болгария и Румыния*
Dativ (D)	*иG=программа# 1АиGФrЧruq= по_программа_расширенный_передовой_сила*
1.3.2 Формализованное представления предикатно-актантной структуры предложений (PSORus)	
PSOL=z5ЎТЊТТЊй*=создать,укрепить;альянс;группировка;страна*	
1.3.3 Формализованное представления «скелета» предложений (SklRus)	
1АиGААЮОэАЮОwS7АиЕяАДqxS7АхSuGxSvGxSэАхS*=по;программа_0865383;альянс_0137657; создать_0987537;и;укрепить_1209752; группировка_04378546;в;страна_1036572# 1АиGФrЧruqЦUААЮОэАЮОФrwS7АиЕшАяАФhДqxS7АхSuGxSvGxSэАхS*	

1.3.4 Формализованное представления предложений (SenRus)	
1AuGФrЧruqЦUAAЮОэAЮОФrwS7AuEшAяAФhДqxS7AxSuGxSvGxSэAxS=*по_программа_ расширенный_передовой_сила_североатлантический_альянс_создать_и_укрепить_военный_ группировка__ страна ЕС вдоль западный граница Россия: в Латвия, Литва, Эстония Польша, Болгария и Румыния*	
Исходное предложение №4 инфоповода-IE-2671-173_07.12.19	
В каждом из государств Восточной Европы разместят около полутора тысяч военных и сотни танков, артиллерийских установок и прочей военной техники	
1.4.1 Формализованное описание элементов предикатно-актантной структуры предложения -IE-2671_173_4-07.12.19 Type=POL={V,N,FN}={разместить;войско;страна}	
Predicate (P)	*гШ =разместить# гШ = разместить*
Subject (S)	*-*
Object (O)	*Фs,FA,wB,wB=военный, танк, установка, техника# яАщAvBФsэАГЦFАЦUwBэA8FФswB =около полутора тысяча военный и сотня танк, артиллерийский установка и прочая военный техника*
Location (L)	*ЖА =государство # 7АФіяАЖАФhuq = в каждый из государство восточный европа*
1.4.2 Формализованное представления предикатно-актантной структуры предложений (PSORus)	
POL=z5ЎTЊTTЊй*=разместить;военный,танк,установка,техника;государство*	
1.4.3 Формализованное представления «скелета» предложений (SklRus)	
TЊTЎTЊDЊA€TЊй=*в;государство_0465783;разместить_0876493;около;военный_0287453;и;танк_114562;установка_1328793;и;техника_1198453# 3TЊйTz5Ў4DTЊQ3ADЊDA€vTDЊййbzн€DAЊ3A*	
1.4.4 Формализованное представления предложений (SenRus)	
3TЊйTz5Ў4DTЊQ3ADЊDA€vTDЊййbzн€DAЊ3A=*в_каждый_из_государство_восточный_европа_ разместить_ около_полутора_тысяча_военный_и_сотня_танк,_артиллерийский_установка_ и прочая военный техника*	
Исходное предложение №5 инфоповода-IE-2671-173_07.12.19	
Кроме того, в Черном море НАТО создало минигруппу флота, в которую (минигруппу) входят не только причерноморские страны, но еще и Польша, Испания и Германия.	
1.5.1.1 Формализованное описание элементов предикатно-актантной структуры предложения -IE-2671_173_1-07.12.19 Type=PSOL={V,N,N,FN}={создать;нато;войско;море}	
Predicate (P)	*dd =создать# dd = создать*
Subject (S)	*щA =нато# щA = нато*
Object (O)	*AB =минигруппа# unAB = минигруппа флот*
Location (L)	*ИG=море# 7АФiИG = в черный море*
1.5.1.2 Формализованное описание элементов предикатно-актантной структуры предложения -IE-2671_173_1-07.12.19 Type=POL={V,N,FN}={входить;войско;страна нато}	
Predicate (P)	*гШ=входить# гШ=входить*
Subject (S)	*-*
Object (O)	*AB =минигруппа# unAB = минигруппа флот*
Location (L)	*uq=страна#ЦTuqэAюAэAvBxЦэAxЦ=причерноморский_страна,_но_еще_и_Польша,_ Испания и Германия*
1.5.2 Формализованное представления предикатно-актантной структуры предложений (PSORus)	
ddunAB7АФiИG *= создать; минигруппа;море* ***ddun7АИG*** *= входить; минигруппа;страна*	
1.5.3 Формализованное представления «скелета» предложений (SklRus)	
7АИGщAddun= *в; море_076784;нато_0658341;создать_09875374; минигруппа_076784; входить_076784;страна_076784=3TЊйTz5Ў4DTЊQ3ADЊDA€vTDЊййbzн€DAЊ3A*	
1.5.4 Формализованное представления предложений (SenRus)	
3TЊйTz5Ў4DTЊQ3ADЊDA€vTDЊййbzн€DAЊ3A=*кроме_того,_в_черный_море_нато_создать_ минигруппа_флот_в_который_входить_не_только_причерноморский_страна,_но_еще_и_Польша,_ Испания и Германия*	

Примечание: В заголовке каждого формализованного описания предложения указывается идентификатор события и приводится его исходная текстовая форма. В разделах 1.1.1, 1.3.1, 1.4.1, 1.5.1 представлен тип предикатно-актантной структуры, его состав (в мнемонике падежной грамматики и грамматических классов слов), унифицированный «скелет», идентификатор события, номер предложения, а также приводятся все элементы предикатно-актантной структуры в сокращенной и полной форме.

В разделах 1.1.2, 1.3.2, 1.4.2, 1.5.2 дано формализованное представление предикатно-актантной структуры (*SPORus*) предложения в сокращенной форме в виде символов обобщенных синтагм и в виде унифицированных форм главных слов предикатно-актантной структуры. Входом в этот раздел служит усеченная поисковая синтагма предикатно-актантной структуры.

В разделах 1.1.3, 1.3.3, 1.4.3, 1.5.3 приводится формализованное представления «скелета» (*SklRus*) предложений события (в сокращенном виде – только главные слова, в полном виде эти элементы сопровождаются номерами наименований понятий эталонного концептуального словаря). Входом в этот раздел служит усеченная поисковая синтагма «скелета» предложения, а замыкает раздел поисковая синтагма предложения.

В разделах 1.1.4, 1.3.4, 1.4.4, 1.5.4 можно найти формализованное представление предложений (*SenRus*) в виде последовательности нормальных слов. Входом в этот раздел служит поисковая синтагма предложения.

Таблица 8

Формализованное и индексное представление элементов предикатно-актантной структуры события (*IE-2671_173_1-07.12.19)*

Формализованное представления элементов предикатно-актантной структуры события
PSOLD={V,N,N,FN}-> ***P****={перебрасывать_**044**_0865383,создать_**054**_0987537,укрепить_**057**_1209752 ,разместить_**047**_0876493}; **S**={нато_**038**_0658341}; **O**={войско_**027**_0378327}; **L**={страна_**022**_09334562,море_**021**_076784}; **D**={программа_**016**_0865383}=**000237964***
Индексное представление элементов предикатно-актантной стуктуры события
перебрасывать_{**044**_0865383_P_000237964}, создать_{**054**_0987537_P_000237964}, укрепить_{**057**_1209752_P_000237964}, разместить_{**047**_0876493_P_000237964}, нато_{**038**_0658341_S_000237964}, войск_{**027**_0378327O_000237964}, страна _{**022**_0865383_L_ 000237964}, море_{**021**_09334562_L_000237964},программа_{**016**_0865383_D_000237964}

Для каждого предложения в табл. 7 указываются различные формы представления единиц смысла в виде их синтаксических конструкций: 1) формализованное представления элементов предикатно-актантной структуры предложений; 2) формализованное представление предикатно-актантной структуры предложения; 3) формализованное представления «скелета» предложений; 4) формализованное представления предложений. Каждая синтаксическая конструкция представляется в сокращенном виде (в виде главного слова конструкции) и полной форме.

Для генерации формализованного представления обобщенного смыслового содержания события суммировались полные формы представления элементов предикатно-актантной структуры предложений. Результаты приведены в табл. 8 в виде формализованного и индексного представления элементов предикатно-актантной структуры события, каждый из которых сопровождается весовым коэффициентом, индексом синтаксической роли, номером словосочетания в эталонном концептуальном словаре и уникальным идентификатором события.

В этих представлениях номер события был заменен на его уникальный идентификационный номер (***ID=000237964***)

Таким образом, как видно из табл. 7 и 8, основная задача формализации смыслового представления описания события заключается в генерации совокупности унифицированных представлений предикатно-актантной структуры. Полученная совокупность формализованных представлений, «скелета» и предложения не только дает возможность поиска по любому элементу их формализованного описаний, но и обеспечивает последовательный переход от каждого нижестоящего элемента формализованного описания к вышестоящему, а также переход в обратном порядке. Такое представление смысла предложений позволяет реализовать весь спектр семантических операций над смысловым содержанием событий.

Методы и алгоритмы формализации синтаксической структуры конструкций исходного предложения и преобразования их в поисковые представления конструкций предложений подробно изложены в работе [6, 9].

[4] Структура предложения в виде главных слов словосочетаний и их отношений.

Полученное формализованное представление обобщенного смыслового содержания событий обеспечивает как поиск и сопоставление идентичных или близких событий, так и возможность их классификации по различным основаниям: по содержанию, по именам субъектов или объектов событий или по тональности отношений между ними и др.

КЛАССИФИКАЦИЯ ИНФОРМАЦИОННЫХ СОБЫТИЙ

Существующие методы классификации текстовых документов можно разделить на несколько основных типов: 1) вероятностные методы классификации [11, 12]; 2) методы, базирующиеся на машинном обучении [13, 14]; 3) методы, основанные на категоризации знаний ("инженерный подход") [5].

При ориентации на вероятностные методы построения классификаторов спользуются предварительно отрубрицированные коллекции документов. Применение методов машинного обучения предполагает предварительную ручную разметку коллекции документов. При применении методов, основанных на знаниях, правила отнесения документа к той или иной рубрике задаются экспертами на основе анализа содержания рубрикатора и содержания документов предметной области. Принципиальная разница между этими методами состоит в том, что при вероятностном методе и методе, основанном на машинном обучении, применяют математические модели для извлечения знаний на основе выявленных закономерностей в обучающей коллекции текстов, в то время как "инженерный подход" использует знания эксперта, основанные на его представлении о содержании документов и правилах отнесения документов к рубрикам классификатора.

Необходимо отметить, что для всех этих методов классификации характерна значительная трудоемкость и слабая ориентация на анализ смысловой структуры классифицируемых текстов. Между тем, в рамках наших исследований процесс классификации текстов был ориентирован именно на смысловой анализ. Этот процесс можно представить как сопоставление формализованных смысловых описаний текстов (представленных в терминах обобщенных наименований понятий и их смысловых отношений) с формализованным смысловым описанием содержания

рубрик классификатора. Решение задачи классификации текстовых описаний информационных событий должно быть ориентировано на автоматизированное создание классификаторов и разработку механизмов сопоставления формализованной смысловой структуры информационных событий и формализованного представления содержания рубрик классификатора.

При создании классификаторов должна быть также разработана технология автоматизированного создания классификационных словарей, в которых отражается в обобщенном виде понятийный состав содержания рубрик классификатора. Каждое понятие сопровождается весовым коэффициентом, характеризующим меру значимости наименования понятия в рубрике. Эти словари можно автоматически сформировать на основе анализа понятийного состава заранее отрубрицированных описаний информационных событий. При таком анализе для каждого классификационного понятия возможно указать его синтаксическую роль в предложениях события.

Установление степени смысловой близости описаний информационных событий и рубрик классификатора можно вычислить методом установления угла косинуса между векторами значений мер наименований понятий по следующей формуле:

$$\mu\left(n_i, n_j\right) = \frac{\sum_t n_{it} \cdot n_{jt}}{\sqrt{\sum_m n_{im}^2} \cdot \sqrt{\sum_k n_{jk}^2}} , \qquad (5)$$

где: n_{it} – значение веса наименования понятия t в событии

n_{jt} – значение веса наименования понятия t в рубрике классификатора

$\sum_m n_{im}$ – сумма всех значений наименований понятий в событии n_i;

$\sum_k n_{jk}$ – сумма всех значений наименований понятий в рубрике классификатора n_j.

Дополнительно для повышения точности классификации событий можно использовать степень соответствия полноты представлений их смысловой структуры и содержания рубрик. В табл. 9 приведены типы соответствий смысловых структур и соответствующих им коэффициентов полноты представлений смысловых структур событий.

С учетом дополнительных условий, связанных с вычислением меры полноты соответствия смысловых структур событий и рубрик, уточненная мера степени смысловой близости выявленных в текстах СМИ событий и рубрик классификатора вычисляется по формуле:

$$M_s = \mu\left(n_i, n_j\right) \cdot K_p , \qquad (6)$$

где: $p = \{1,2,3,4,5,6\}$.

При использовании предлагаемой модели представления смыслового содержания документов структура классификатора может быть представлена в виде перечня унифицированных наименований понятий, сопровождаемых их весовым коэффициентом, индексом синтаксической роли и номером рубрики (табл. 10). В случаях, когда одно и то же понятие принадлежит к разным рубрикам, указываются все рубрики с их атрибутами (весом, индексом роли и номером рубрики).

Таблица 9

Таблица коэффициентов полноты представлений смысловых структур событий

№ п/п	*Тип соответствия смысловой структуры*	*Коэффициент полноты представления смысловых структур событий*
01	***PSOLD={V,N,N,FN}***	***K1=7***
02	***PSO={V,N,N,FN}***	***K2=6***
03	***PSOD={V,N,N,FN}***	***K3=5***
04	***PSO={V,N,N}***	***K4=4***
05	***PS={V,N}***	***K5=3***
06	***PO={V,N}***	***K6=2***

Таблица 10

Предлагаемая структура классификатора

нато_{038_S_023},{032_O_214},{037_S_947}
войск_{027_O_ 023},{034_O_214}
перебрасывать_{044_P_023},{061_P_947}
создать_{054_P_023,{067_P_023}
страна_{022_L_023},{041_O_023},{034_L_346},{027_L_947}
разместить_{047_P_023},{057_P_214}
море_{021_L_ 023}
программа _{016_D_023},{047_O_214}
укрепить_{057_P_023},{063_P_214}

Исходя из вышеизложенного, процесс классификации событий можно представить как сопоставление элементов их индексного представления и элементов индексного представления рубрик классификатора и вычисления меры смысловой близости описаний событий и рубрик классификатора по формулам (5) и (6).

ЗАКЛЮЧЕНИЕ

В настоящей статье предлагается решение проблемы автоматического выявления, формализации и классификации информационных событий в текстах СМИ, базирующееся на разработанных нами методах и алгоритмах автоматического выявления описаний информационных событий и методах преобразования текстового представления событий в их унифицированное смысловое представление, позволяющее в значительной степени решить проблемы вариативности представлений смыслового содержания событий. Новизна предлагаемого решения обусловлена тем фактом, что в основу этих методов положена уникальная машинная грамматика [8] и семантико-синтаксический анализ текстов [6], тесно увязанный с реализацией правил падежной грамматики Ч. Филлмора. Индексное обобщенное представление содержания документов и индексное представление структуры классификатора позволяет обеспечить высокую эффективность процедур идентификации, кластеризации и классификации информационных событий в текстовом потоке новостной информации. Проведенные эксперименты на массиве сообщений СМИ показали работоспособность предлагаемых решений.

СПИСОК ЛИТЕРАТУРЫ

1. Богатырев М.Ю. Извлечение фактов из текстов естественного языка с применением концептуальных графовых моделей // Известия ТулГУ. Технические науки. – 2016. – № 7, Ч. 1. – С. 198-207
2. Виноградов А.Н., Власова Н.А., Куршев Е.П., Подобряев А.В. Современные технологии обработки естественного языка в задачах стратегического управления // Технологическая перспектива в рамках евразийского пространства: новые рынки и точки экономического роста. – СПб: Центр научно-информационных технологий "Астерион", 2018.
3. Ермаков А.Е. Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей // Компьютерная лингвистика и интеллектуальные технологии : труды Международной конференции «Диалог'2007». – М.: Наука, 2007.
4. Хорошилов Ал-др.А., Никитин Ю.В., Хорошилов Ал-ей.А., Будзко В.И. Автоматическое создание формализованного представления смыслового содержания неструктурированных текстовых сообщений СМИ и социальных сетей // Системы высокой доступности. – 2014. – Т.10, № 3.
5. Helbig H. Knowledge representation and the semantics of natural language. – Berlin: Springer, 2006.
6. Кан А.В., Ревина В. Д., Руснак В.И., Хорошилов Ал-др А., Хорошилов А.А. Автоматическое формирование синтаксической модели языка для задач машинного перевода и информационного поиска // Научно-техническая информация. Сер. 2. – 2018. – № 12 – С. 25-41.
7. Филлмор Ч. Дело о падеже // Новое в зарубежной лингвистике. Вып. 10. – М.: Прогресс, 1981.
8. Аблов И.В., Козичев В.Н., Ширманов А.В., Хорошилов Ал-др А., Хорошилов Ал-ей А. Средства машинной грамматики русского языка (по Г.Г. Белоногову) // Научно-техническая информация. Сер. 2. – 2018. – № 6. – С. 32-46.
9. Калинин Ю.П., Хорошилов Ал-др А., Хорошилов Ал-ей А. Современные технологии автоматизированной обработки текстовой информации // Системы высокой доступности. – 2015. – Т.11, № 2. – С. 67-79.
10. Захаров В.Н., Мусабаев Р.Р., А.М. Красовицкий А.М., Козловская Я.Д., Хорошилов Ал-др А., Хорошилов Ал-ей А. Метод кластеризации новостных сообщений СМИ на основе их концептуального анализа // Информатика и её применение. – 2019. – Т. 29, № 3. – С. 52-65
11. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: классификация и снижение размерности. – М.: Финансы и статистика, 1989.
12. Алон Н., Спенсер Дж. Вероятностный метод. – М.: Бином, 2007.
13. Гольдберг Й. Нейросетевые методы в обработке естественного языка. – М.: ДМК, 2019.
14. Осинга Д. Глубокое обучение. Готовые решения. – СПб: Диалектика, 2019.

Материал поступил в редакцию 21.04.2020

Сведения об авторах

ХОРОШИЛОВ Александр Алексеевич – доктор технических наук, профессор Московского авиационного института (Национальный исследовательский университет) (НИУ МАИ), ведущий научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН (ФИЦ ИУ РАН), старший научный сотрудник 27-го Центрального научно-исследовательского института Министерства обороны Российской Федерации (27 ЦНИИ МО РФ), Москва
e-mail: khoroshilov@mail.ru

МУСАБАЕВ Рустам Рафикович – кандидат технических наук, руководитель лаборатории Института информационных и вычислительных технологий, Республика Казахстан
e-mail: rmusab@gmail.com

КОЗЛОВСКАЯ Яна Дмитриевна – студент НИУ МАИ, Москва
yana04029877@mail.ru

НИКИТИН Юрий Викторович – научный сотрудник ФИЦ ИУ РАН
e-mail: yuri.v.nikitin@gmail.com

ХОРОШИЛОВ Алексей Александрович – кандидат технических наук, научный сотрудник ФИЦ ИУ РАН
e-mail: a.a.horoshilov@mail.ru

Ирина Владимировна Маршакова
(1941–2019)

Ушла из жизни доктор философских наук Ирина Владимировна Маршакова – известный специалист по наукометрии, она принадлежит к первому поколению отечественных науковедов, исследовавших цитируемость научных публикаций в 1960-е гг., когда это направление науки только формировалось, а термин «наукометрия» еще не был введен В.В. Налимовым, который проводил эти работы в МГУ им М.В. Ломоносова.

И.В. Маршакова внесла большой вклад в развитие наукометрии, создав одновременно с сотрудником Института научной информации США Г. Смоллом (*Henry Small, ISI*) и независимо от него кластеры социтирования на массиве статей по лазерам. Будучи аспиранткой Ю.А. Шрейдера в ВИНИТИ РАН, она опубликовала результаты этого исследования в статье «Система связей между документами, построенная на основе ссылок (по указателю *Science Citation Index*)» (НТИ. Сер. 2. – 1973. – № 6. – С. 3-8.), которая была высоко оценена Ю. Гарфилдом и к настоящему времени процитирована в *WoS* 129 раз**.**

В 1975 г. И.В. Маршакова защитила кандидатскую диссертацию «Алгоритмические классификации динамических документальных массивов информации» в ВИНИТИ РАН, а в 1993 г. – докторскую «Методы количественного анализа научного знания» в Институте философии РАН, где до конца жизни работала ведущим специалистом. Она участвовала в первой международной конференции в Берлине в 1993 г., на которой было организовано Международное общество по информетрии и наукометрии (*International Society for Informetrics and Scientometrics- ISSI*), и в международной конференции там же в 2000 г., на которой было организовано Международное общество научного сотрудничества по наукометрии (*Collaborative Network – COLLNET*), активно принимала участие во многих конференциях этих организаций, выступая с научными докладами. Она пользовалась заслуженным авторитетом у коллег по цеху.

Круг интересов Ирины Владимировны был обширен: история и философия науки и ее библиометрический анализ, информационные технологии, специфика использования библиометрических показателей для оценки гуманитарных направлений научного знания. Продолжая исследования по библиометрии, она вела большую преподавательскую работу в университетах Польши: Силезском университете в Катовицах (*Uniwersytet Śląski w Katowicach*), а затем в университете Казимира Великого в Быдгоще (*Uniwersytet Kazimierza Wielkiego w Bydgoszczy*) в 1995–2000 гг., руководила научным проектом, поддержанным Фондом Дж. Сороса, и многими отечественными проектами, поддержанными РГНФ и РФФИ. В последние годы жизни сферой её научных интересов были нормализованные показатели оценки значимости научных журналов.

Мы всегда ощущали участие Ирины Владимировны в деятельности наукометрического и информационного сообщества, и ее внезапный уход нанес ощутимый урон нашей науке, специалисты которой будут всегда помнить эту умную и красивую женщину.

Редколлегия и редакция сборника
«Научно-техническая информация»

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

ВИНИТИ РАН, как единственный в России владелец лицензии Консорциума УДК, предлагает издания УДК полного четвертого издания на русском языке в печатном и электронном виде:

1. Таблицы УДК

УДК. Том I Общая методика применения УДК. Вспомогательные таблицы. Основные таблицы. Общий отдел. Алфавитно-предметный указатель к Общему отделу

УДК. Том II **1/3** Философия. Психология. Религия. Богословие. Общественные науки (только электронное издание)

УДК. Том III **5/54** Математика. Естественные науки (только электронное издание)

УДК. Том IV **55/59** Геологические и биологические науки (только электронное издание)

УДК. Том V **6/61** Медицинские науки (только электронное издание)

УДК. Том VI (часть 1) 6/621 Прикладные науки. Технология. Инженерное дело (только электронное издание)

УДК. Том VI (часть 2) 622/629 Техника. Инженерное дело (только электронное издание)

УДК. Алфавитно-предметный указатель к т. VI (1 и 2 части) (только электронное издание)

УДК. Том VII **63/65** Сельское хозяйство. Домоводство. Управление предприятием (только электронное издание)

УДК. Том VIII **66** Химическая технология. Химическая промышленность. Пищевая промышленность. Металлургия. Родственные отрасли (только электронное издание)

УДК. Том IX **67/69** Различные отрасли промышленности и ремесел. Строительство (только электронное издание)

УДК. Том X **7/9** Искусство. Спорт. Филология. География. История.

УДК. АПУ (с в о д н ы й) к полному 4-му изданию

УДК. Изменения и дополнения. Выпуск 2 (к т.т. 1–3) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 3 (к т.т. 1–6) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 4 (к т.т. 1–7) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 5 (к т.т. 1–10)

УДК. Изменения и дополнения. Выпуск 6 (к т.т. 1–10)

УДК. Изменения и дополнения. Выпуск 7 (к т.т. 1–10), 2017 г. (только электронное издание)

Для подписки необходимо направить заявку по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефоны: 499-155-42-85, 499-151-78-61

E-mail: feo@viniti.ru